A Posteriori Error Estimation for Elliptic Homogenization Problems

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde

(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Stephanie Meier-Rohr

von Staufen AG

Promotionskommission

Prof. Dr. Stefan A. Sauter (Leitung der Dissertation)

Prof. Dr. Rémi Abgrall

Prof. Dr. Michel Chipot

Zürich, 2018

Abstract

In the simulation of complicated physical phenomena, errors due to mathematical modelling and numerical discretization arise. A priori and a posteriori estimates for numerical discretization errors are well studied, while modelling errors are a topic of vivid research. A recent, useful development are functional a posteriori error majorants, which are of importance for the reliability of the solution. In this thesis, we consider the development of an a posteriori error estimate for the modelling and discretization error for periodic structures as they appear, e.g., in the design of composite materials, where the behaviour is modelled by homogenization. Further, an adaptive algorithm which controls both error terms is developed.

Homogenization theory is well developed in the literature and was originally introduced for the study of composite material. Such type of materials consist of a main homogeneous material, with small heterogeneities, that can be modelled as a periodic structure. To solve a boundary value problem on a periodic structure, one has to resolve the small heterogeneities, which is numerically not feasible. Therefore, by thinking of two scales, one solves microscopic cell problems and a macroscopic homogenized problem. With this knowledge, we can construct a two scale approximation of the original problem.

This fully discrete solution is not a Galerkin approximation of the original boundary value problem. Hence, for developing a posteriori error estimates, majorants of functional type are a suitable approach. We present a fully computable total error majorant, consisting of the majorant for the cell problems, the majorant for the homogenized problem and a third term related to the two scale approximation error. The knowledge of these different error contributions allows us to develop an error estimation strategy. Moreover, we study a suitable gradient recovery procedure, as it is needed for the majorants.

In the numerical experiments, we discuss the efficiency of the gradient recovery procedure and the sharpness of the majorants. The behaviour of the total error and majorant depending on the different scales is interesting to observe.

Zusammenfassung

Bei der Simulation von komplizierten physikalischen Phänomenen treten Fehler sowohl durch mathematische Modellierung als auch durch numerische Diskretisierung auf. Während a priori und a posteriori Abschätzungen für numerische Diskretisierungsfehler weit entwickelt sind, wurden Modellierungsfehler erst in den letzten Jahren untersucht. Eine neue Entwicklung stellen dabei funktionale a posteriori Fehlermajoranten dar, welche von grosser Bedeutung für die Zuverlässigkeit der Lösung sind. Diese Doktorarbeit behandelt die Entwicklung von a posteriori Fehlerschätzern für Modellierungs- und Diskretisierungsfehler periodischer Strukturen, wie sie z.B. in der Entwicklung von Verbundwerkstoffen auftauchen, wobei das Verhalten mittels Homogenisierung modelliert wird. Zudem wird ein adaptiver Algorithmus entwickelt, welcher die beiden Fehlerterme kontrolliert.
Homogenisierung ist in der Literatur ausführlich behandelt und wurde ursprünglich für die Entwicklung von heterogenen Materialien eingeführt. Materialien dieser Art bestehen aus einer homogenen Grundsubstanz, mit kleinen heterogenen Komponenten, welche als periodische Struktur modelliert werden können. Um ein Randwertproblem auf einer periodischen Struktur zu lösen, muss man die kleinen heterogenen Strukturen auflösen, was numerisch zu aufwändig ist. Deshalb arbeitet man mit einem Zwei-Skalen-Modell und löst ein mikroskopisches Zellenproblem und ein makroskopisches homogenisiertes Problem. Mit diesem Wissen können wir eine Zwei-Skalen-Approximation des ursprünglichen Problems konstruieren.
Diese volldiskrete Lösung ist keine Galerkin Approximation des ursprünglichen Randwertproblems. Folglich sind für die Entwicklung von a posteriori Fehlerschätzern Majoranten von funktionalem Typ der geeignete Ansatz, im Gegensatz zu residualen Fehlerschätzern. Wir präsentieren eine berechenbare Gesamtfehlermajorante, welche aus der Majorante des Zellenproblems, der Majorante des homogenisierten Problems und einem dritten Term besteht, welcher dem Zwei-Skalen-Approximationsfehler entspricht. Die Kenntnis dieser drei unterschiedlichen Fehlerquellen erlaubt uns die Entwicklung einer adaptiven Fehlerabschätzungsstrategie. Zudem untersuchen wir ein passendes Gradienten-Glättungsverfahren, da dieses für die Majoranten benötigt wird.
In den numerischen Experimenten diskutieren wir die Effizienz des Gradienten-Glättungsverfahrens und die Genauigkeit der Majorante. Zudem werden wir das Verhalten des Gesamtfehlers und der Gesamtfehlermajorante für verschiedene Skalen untersuchen.

Acknowledgement

I would like to express my sincere gratitude to my advisor Prof. Dr. Stefan Sauter for his continuous support, fruitful discussions and for giving me the opportunity to do my PhD on such an interesting topic.
Furthermore, I would like to thank Prof. Dr. Sergey Repin for his helpful inputs during my short visit at the University of Jyväskylä.
I am also grateful to Dr. Svetlana Matculevich, who was always interested in a discussion on practical implementation.
Last but not least, I would like to thank my husband, my parents, my sister and my friends for their patience and motivation during this time.

Zurich, September 2017 Stephanie Meier-Rohr

Table of Contents

List of Symbols

Ω	Domain, open set in $\mathbb{R}^d$, p. 92
$\widehat{\Pi}$	Reference cell in $\mathbb{R}^d$, p. 9
$\Pi_{\mathbf{i}}^{\varepsilon}$	General cell in a periodic structure, in $\mathbb{R}^d$, p. 25
$\mathbf{A}_\varepsilon, \mathbf{b}_\varepsilon, c_\varepsilon$	Coefficients on the domain, p. 25
$\widehat{\mathbf{A}}, \widehat{\mathbf{b}}, \hat{c}$	Coefficients on the reference cell, p. 25
u_ε	Solution of the homogenization problem, p. 26
$\widehat{N}_j$	Solution of a cell problem, p. 30
$\mathbf{A}_0, B_0, c_0$	Homogenized coefficients, p. 30
u_0	Solution of the homogenized problem, p. 30
u_ε^1	Two scale approximation, p. 30
w_ε^1	Boundary corrected two scale approximation, p. 30
φ_ε	Cutoff function, p. 30
$\langle\cdot\rangle_\Omega$	Mean value of a function, p. 9
$\alpha_\varepsilon^{\text{cont}}$	Continuity constant of $\mathbf{A}_\varepsilon$, p. 25
$\alpha_\varepsilon^{\text{ell}}$	Ellipticity constant of $\mathbf{A}_\varepsilon$, p. 25
α_0^{ell}	Ellipticity constant of $\mathbf{A}_0$, p. 33
α_0^{cont}	Maximal eigenvalue of $\mathbf{A}_0$, p. 34
$\alpha_{0,l}^{\text{cont}}$	Maximal eigenvalue of $\mathbf{A}_{0,l}$, p. 38
$\alpha_{0,l}^{\text{ell}}$	Minimal eigenvalue of $\mathbf{A}_{0,l}$, p. 38
$\widehat{\alpha}^{\text{cont}}$	Continuity constant of $\widehat{\mathbf{A}}$, p. 25
$\widehat{\alpha}^{\text{ell}}$	Ellipticity constant of $\widehat{\mathbf{A}}$, p. 25
G_h	Gradient recovery operator, p. 21
Q_h	L^2-projection operator, p. 22
S	Smoothing operator, p. 22

$\mathcal{M}$	Majorant, p. 19
$\mathcal{M}_{\mathrm{D}}$	Dual term of the majorant, p. 21
$\mathcal{M}_{\mathrm{Eq}}$	Equilibrium term of the majorant, p. 21
$\mathcal{M}_{\mathrm{disc}}\left(\widehat{N}_k^{(l)}\right)$	Majorant for the cell problem, p. 40
$\mathcal{M}_{\mathrm{disc}}\left(u_0^{(l,j)}\right)$	Majorant for the homogenized problem, p. 42
$\mathcal{M}_{\mathrm{tot}}$	Total error majorant, p. 44
Eff_G	Efficiency index of the gradient recovery procedure, p. 71
$\mathrm{Eff}_{\mathcal{M}}$	Efficiency index of the majorant, p. 74
ψ_j	Finite element basis function, p. 12
$C_{F\Omega}$	Constant from the Friedrichs inequality, p. 97
$C_{P\Omega}$	Constant from the Poincaré inequality, p. 97
C_{PW}	Constant from the Poincaré-Wirtinger inequality, p. 97
$C_{T\Gamma}$	Constant from the trace inequality, p. 97
$\mathbf{C}_h$	Clément operator, p. 98
div	Divergence, p. 92
∇	Gradient, p. 92
Δ	Laplace operator, p. 92
$\mathcal{P}_t$	Set of polynomials of degree $\leq t$, p. 13
$\rho_\Omega(\cdot)$	Spectral radius of a matrix function, p. 91
supp	Support, p. 94
V_h	Discrete finite element space, p. 14
X'	Dual space, p. 93
$L^p(\Omega)$	Real-valued Lebesgue space for $1 \leq p < \infty$, p. 92
$L^\infty(\Omega)$	Real-valued Lebesgue space for $p = \infty$, p. 93
$W^{m,p}(\Omega)$	Real-valued Sobolev space for $m \in \mathbb{N}$ and $1 \leq p \leq \infty$, p. 94
$W_0^{m,p}(\Omega)$	Completion of $C_0^\infty(\Omega)$ with respect to the norm $\|\cdot\|_{W^{m,p}(\Omega)}$, p. 94
$H^m(\Omega)$	Real-valued Sobolev space $W^{m,p}(\Omega)$ for $m \in \mathbb{N}$ and $p = 2$, p. 94

$H(\Omega, \mathrm{div})$	Real-valued Hilbert space, p. 96
$H^1_{\mathrm{per}}(\widehat{\Pi})$	Real-valued Hilbert space of $\widehat{\Pi}$-periodic functions, p. 9
$W_{\mathrm{per}}(\widehat{\Pi})$	Real-valued Sobolev space of $\widehat{\Pi}$-periodic functions, p. 9
$\langle \cdot, \cdot \rangle$	Scalar product, p. 91
$(\cdot, \cdot)_{L^2(\Omega)}$	L^2 scalar product, p. 93
$(\cdot, \cdot)_{H^m(\Omega)}$	H^m scalar product, p. 94
$(\cdot, \cdot)_{\mathrm{div}}$	$H(\Omega, \mathrm{div})$ scalar product, p. 96
$\| \cdot \|$	Vector norm, p. 91
$\| \cdot \|_F$	Frobenius norm, p. 91
$\| \cdot \|_{L^p(\Omega)}$	L^p norm, p. 92
$\| \cdot \|_{L^\infty(\Omega)}$	L^∞ norm, p. 93
$\| \cdot \|_{W^{m,p}(\Omega)}$	$W^{m,p}$ norm, p. 94
$\lvert \cdot \rvert_{W^{m,p}(\Omega)}$	$W^{m,p}$ semi-norm, p. 94
$\| \cdot \|_{\mathbf{A}}$	Energy norm, p. 97
$\| \cdot \|_{\mathbf{A}^{-1}}$	Complementary energy norm, p. 97

1 Introduction

In this thesis, we discuss the homogenization of an elliptic boundary value problem within a periodic structure. Thus, we have a basic cell with microscopic scale $\mathbf{y} = \frac{\mathbf{x}}{\varepsilon}$ and the domain Ω then consists of those repeated and scaled cells with macroscopic variable $\mathbf{x}$. In this setting, the coefficients are periodic and possess further properties, explained later on. For $f \in L^2(\Omega)$, we consider the elliptic partial differential equation of second order

$$-\operatorname{div}(\mathbf{A}_\varepsilon \nabla u_\varepsilon) + \langle \mathbf{b}_\varepsilon, \nabla u_\varepsilon \rangle + c_\varepsilon u_\varepsilon = f \quad \text{in } \Omega,$$

with Dirichlet boundary condition $u_\varepsilon = g$ on $\Gamma = \partial\Omega$, $g \in L^2(\Gamma)$.
For $\varepsilon > 0$ very small, the coefficients $\mathbf{A}_\varepsilon$, $\mathbf{b}_\varepsilon$ and c_ε are rapidly oscillating functions. In the process of finite element discretization, we only get accurate approximations if the mesh size is much smaller than ε, i.e., resolves the oscillations. This problem can be significantly reduced with a two scale approximation

$$u_\varepsilon^1(\mathbf{x}) = u_0(\mathbf{x}) - \varepsilon \sum_{j=1}^{d} \widehat{N}_j(\mathbf{y}) \partial_{x_j} u_0(\mathbf{x}).$$

u_0 is the solution of a homogenized problem on the domain Ω, independent of ε, with the same right-hand side and boundary condition as u_ε and with constant coefficients, i.e., it captures the macroscopic behaviour. $\widehat{N}_j$, for $j = 1, \dots, d$, is the solution of a periodic boundary value problem on the reference cell, with a special right-hand side, i.e., it captures the microscopic behaviour. A priori, one can show the error estimate

$$\left\| u_\varepsilon - u_\varepsilon^1 \right\|_{H^1(\Omega)} \leq c\varepsilon^{\frac{1}{2}},$$

under certain regularity assumptions. Homogenization theory for elliptic partial differential equations is well studied in the literature, see, e.g., [13], [22] and [8].
For the development of a posteriori error estimates, we consider majorants of functional type introduced by [26] and [24]. The advantage of this approach is that no extra regularity or Galerkin orthogonality is required and the majorants do not contain mesh-dependent constants. The goal is to find an error estimator

$$\left\| \nabla \left(u_\varepsilon - u_\varepsilon^1 \right) \right\|_{\mathbf{A}_\varepsilon} \leq \mathcal{M}_{\text{tot}},$$

that is fully computable and gives a guaranteed upper bound. Furthermore, the total error majorant should take into account the error of the cell problems, the error of the homogenized problem and the error due to homogenization.
In [27], a posteriori error majorants are derived for the error purely related to homogenization, i.e., without considering approximation errors of the homogenized and the cell problems. In [28], combined a posteriori modelling/discretization errors are discussed purely related to variable coefficients, without considering the case of periodic coefficients. Alternative approaches include, e.g., the a posteriori estimate of residual type for the heterogeneous multiscale discretizations, cf. [2].

The structure of this thesis is as follows. In Chapter 2, we discuss the variational formulation of a general linear elliptic partial differential equation and study the existence and uniqueness for different boundary conditions, in particular periodic problems. Further, we state fundamental results from the finite element method. Then, we introduce a posteriori error estimates of functional type and the gradient recovery procedure, which is used as a first approximation to compute the majorant.
In Chapter 3, we investigate the theory of homogenization, where we consider the two scale approach. Moreover, we discuss a priori error estimates and study the properties of the homogenized coefficients.
In Chapter 4, we develop the a posteriori error estimate of the two scale approximation. This includes the study of the majorants for the cell and homogenized problems and results in a total

error majorant. First, this is done only with a diffusion matrix, finally, an error estimate for a general reaction-convection-diffusion problem is derived.
In Chapter 5, we explain the implementation procedure. This contains the mesh refinement, assembling of the system matrix and the right-hand side, solving the linear system and further details for special cases. Moreover, we describe the implementation of a general majorant and the gradient recovery procedure. Finally, we mention how to build the two scale approximation.
In Chapter 6, we present several numerical experiments on the choice of parameters and the behaviour of the majorants is shown. Then, we apply the total error majorant to two different homogenization problems.
Finally, we draw the conclusion of this thesis in Chapter 7.
In Appendix A, we give further details about important inequalities and Sobolev spaces.

2 Elliptic Partial Differential Equations

In this chapter we will state important results from the theory of elliptic partial differential equations, as explained in detail, e.g., in [9], [21] and [12]. In particular, we will give the general problem formulation and existence and uniqueness results for different boundary conditions, which we will use in subsequent chapters. Further, we discuss the finite element method and a priori error estimates. Then, we introduce a posteriori error estimates of functional type and the gradient recovery procedure, which will be developed further in Chapter 4.

2.1 Variational Formulation

Let $\Omega \subset \mathbb{R}^d$ be an open, bounded domain with Lipschitz boundary $\Gamma := \partial\Omega$. We consider general linear elliptic partial differential equations of second order in d variables $\mathbf{x} = (x_1, x_2, \ldots, x_d)$. The classical formulation of this problem states: Find $u \in C^2(\Omega) \cap C^0(\overline{\Omega})$ such that

$$\begin{aligned} -\operatorname{div}(\mathbf{A}(\mathbf{x})\nabla u(\mathbf{x})) + \langle \mathbf{b}(\mathbf{x}), \nabla u(\mathbf{x})\rangle + c(\mathbf{x})\,u(\mathbf{x}) &= f(\mathbf{x}) && \text{for } \mathbf{x} \in \Omega, \\ u(\mathbf{x}) &= 0 && \text{for } \mathbf{x} \in \Gamma. \end{aligned} \tag{2.1}$$

The following conditions have to be satisfied: $f \in C^0(\Omega)$, $\mathbf{A} \in C^1(\Omega, \mathbb{R}^{d\times d}_{\mathrm{sym}})$, $\mathbf{b} \in C^0(\Omega, \mathbb{R}^d)$ and $c \in C^0(\Omega, \mathbb{R}_{\geq 0})$. Further, $\mathbf{A}$ is **uniformly elliptic**, i.e., the minimum and maximum eigenvalues $\lambda(\mathbf{x})$, $\Lambda(\mathbf{x})$ of $\mathbf{A}(\mathbf{x})$, fulfil

$$0 < \lambda(\mathbf{x})\,\|\xi\|_2^2 \leq \langle \mathbf{A}(\mathbf{x})\xi, \xi\rangle \leq \Lambda(\mathbf{x})\,\|\xi\|_2^2, \quad \forall \xi \in \mathbb{R}^d\backslash\{0\}. \tag{2.2}$$

The boundary condition $u = 0$ on Γ is called homogeneous Dirichlet boundary condition; we will consider more general boundary conditions in the next section.
By multiplying the first equation of problem (2.1) with an arbitrary function $v \in C_0^\infty(\Omega)$ and building the L^2-scalar product we get:

$$\begin{aligned} (-\operatorname{div}(\mathbf{A}\nabla u) + \langle \mathbf{b}, \nabla u\rangle + cu, v)_{L^2(\Omega)} &= \int_\Omega \langle \mathbf{A}\nabla u, \nabla v\rangle - \int_\Gamma \langle \mathbf{A}\nabla u, \mathbf{n}\rangle\, v + \int \left(\langle \mathbf{b}, \nabla u\rangle\, v + cuv\right) \\ &= (f, v)_{L^2(\Omega)}, \end{aligned}$$

where we used identity (A.4) and the Gaussian Integral Theorem A.1. Since $v \in C_0^\infty(\Omega)$, the boundary integral vanishes. Moreover, we know that $C_0^\infty(\Omega)$ is dense in $H_0^1(\Omega)$, thus we can state the variational formulation: Find $u \in H_0^1(\Omega)$ such that

$$\int_\Omega \left(\langle \mathbf{A}\nabla u, \nabla v\rangle + \langle \mathbf{b}, \nabla u\rangle\, v + cuv\right) = \int_\Omega fv, \quad \forall v \in H_0^1(\Omega). \tag{2.3}$$

We call $u \in H_0^1(\Omega)$ that fulfils (2.3) a weak solution. For the variational formulation the following weaker conditions have to be satisfied:

$$f \in L^2(\Omega), \quad \mathbf{A} \in L^\infty(\Omega, \mathbb{R}^{d\times d}_{\mathrm{sym}}), \quad \mathbf{b} \in L^\infty(\Omega, \mathbb{R}^d) \quad \text{and} \quad c \in L^\infty(\Omega, \mathbb{R}_{\geq 0}). \tag{2.4}$$

We assume $\mathbf{A}$ to be uniformly elliptic.
The variational formulation of the boundary value problem can be generalized in the following way:

Definition 2.1 (Bounded and elliptic bilinear form). *Let H be a Hilbert space. The bilinear form $a : H \times H \to \mathbb{R}$ is called bounded (or continuous), if there exists a constant α^{cont} such that*

$$|a(u, v)| \leq \alpha^{\mathrm{cont}} \|u\|_H \|v\|_H, \quad \forall u, v \in H, \tag{2.5}$$

and H-elliptic, if there exists a constant $\alpha^{\mathrm{ell}} > 0$ such that

$$a(u, u) \geq \alpha^{\mathrm{ell}} \|u\|_H^2, \quad \forall u \in H. \tag{2.6}$$

Theorem 2.2 (Lax-Milgram). *Let a be a bounded, H-elliptic bilinear form on H and $l : H \to \mathbb{R}$ a linear form. Then, the variational formulation: find $u \in H$ such that*

$$a(u, v) = l(v), \quad \forall v \in H,$$

has a unique solution. Moreover, it holds

$$\|u\|_H \leq \frac{1}{\alpha^{\text{ell}}} \|l\|_{H'}.$$

For the variational formulation (2.3) it is clear that we have the bilinear form

$$a(u, v) := \int_\Omega \left(\langle \mathbf{A} \nabla u, \nabla v \rangle + \langle \mathbf{b}, \nabla u \rangle v + cuv \right) \tag{2.7}$$

and the linear form

$$l(v) := \int_\Omega fv. \tag{2.8}$$

To ensure a unique solution for different boundary conditions, we have to verify the conditions of the Lax-Milgram Theorem 2.2 in each case, which will be subject of the next section.
Note that we consider only real-valued function spaces, hence we always treat real-valued bilinear and linear forms.

2.2 Boundary Conditions

Below we will study the following boundary conditions:

a) Homogeneous Dirichlet boundary condition $u = 0$ on Γ, in Subsection 2.2.1.

b) Inhomogeneous Dirichlet boundary condition $u = g$ on Γ, in Subsection 2.2.1.

c) Neumann boundary condition $\langle \mathbf{A} \nabla u, \mathbf{n} \rangle = g$ on Γ, in Subsection 2.2.2.

d) Periodic boundary condition u $\widehat{\Pi}$-periodic, in Subsection 2.2.3.

The periodic boundary condition plays an essential role in the homogenization theory as examined in Chapter 3. In this section, we will always assume that (2.4) holds.

2.2.1 Dirichlet Boundary Condition

For the homogeneous Dirichlet boundary condition, i.e. $u = 0$ on Γ, we have already derived the bilinear form (2.7) and the linear form (2.8). Now, we prove the unique solvability:

Proposition 2.3. *Assume that* (2.4) *is fulfilled,* $\mathbf{A}$ *satisfies* (2.2), $\operatorname{div} \mathbf{b} \in L^\infty(\Omega, \mathbb{R})$ *and* $-\frac{1}{2} \operatorname{div} \mathbf{b} + c \geq 0$. *Then, the homogeneous Dirichlet problem with variational formulation* (2.3) *has a unique weak solution.*

Proof. In order to prove the statement, we have to verify the conditions of the Lax-Milgram Theorem 2.2. In this case we consider $H := H_0^1(\Omega)$.
With the Cauchy-Schwarz inequality it follows that the bilinear form is bounded:

$$\begin{aligned}
|a(u, v)| &= \left| \int_\Omega \left(\langle \mathbf{A} \nabla u, \nabla v \rangle + \langle \mathbf{b}, \nabla u \rangle v + cuv \right) \right| \\
&\leq \|\mathbf{A}\|_{L^\infty(\Omega)} \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} + \|\mathbf{b}\|_{L^\infty(\Omega)} \|\nabla u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} + \|c\|_{L^\infty(\Omega)} \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \\
&\leq \Lambda |u|_{H^1(\Omega)} |v|_{H^1(\Omega)} + \|c\|_{L^\infty(\Omega)} \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} + \|\mathbf{b}\|_{L^\infty(\Omega)} \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)} \\
&\leq \left(\max \left\{ \Lambda, \|c\|_{L^\infty(\Omega)} \right\} + \|\mathbf{b}\|_{L^\infty(\Omega)} \right) \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)}.
\end{aligned}$$

Further, the bilinear form is $H_0^1(\Omega)$-elliptic: First, with identity (A.4) and the Gaussian Integral Theorem A.1, we prove the following equality for $u \in H_0^1(\Omega)$:

$$\begin{aligned}\int_\Omega \langle \mathbf{b}, u\nabla u\rangle &= \int_\Omega \frac{1}{2}\left\langle \mathbf{b}, \nabla u^2\right\rangle \\ &= \int_\Omega \frac{1}{2}\operatorname{div}\left(\mathbf{b}u^2\right) - \int_\Omega \frac{1}{2}\operatorname{div}(\mathbf{b})\,u^2 \\ &= \frac{1}{2}\int_\Gamma \left\langle \mathbf{b}u^2, \mathbf{n}\right\rangle \mathrm{d}s - \int_\Omega \frac{1}{2}\operatorname{div}(\mathbf{b})\,u^2 \\ &= -\int_\Omega \frac{1}{2}\operatorname{div}(\mathbf{b})\,u^2.\end{aligned}$$

Then, it follows

$$\begin{aligned}a(u,u) &= \int_\Omega \left(\langle \mathbf{A}\nabla u, \nabla u\rangle + \langle \mathbf{b}, \nabla u\rangle\, u + cu^2\right) \\ &\geq \lambda\|\nabla u\|^2_{L^2(\Omega)} + \int_\Omega \left(c - \frac{1}{2}\operatorname{div}(\mathbf{b})\right)u^2 \\ &\geq \lambda|u|^2_{H^1(\Omega)} \\ &\geq \lambda\frac{1}{1+C_{F\Omega}^2}\|u\|^2_{H^1(\Omega)},\end{aligned}$$

where we used $c - \frac{1}{2}\operatorname{div}(\mathbf{b}) \geq 0$ and Friedrichs inequality (see Section A.3).
Finally, l is a linear form $l : H^1(\Omega) \to \mathbb{R}$:

$$|l(v)| = \left|\int_\Omega fv\right| \leq \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)}\|v\|_{H^1(\Omega)}.$$

Thus, we can apply the Lax-Milgram Theorem 2.2, which gives us existence and uniqueness of the solution. □

The inhomogeneous case, for $u = g$ on Γ, can be treated as follows. First, note that with the Trace Theorem A.23 the restriction $u|_\Gamma$ of $u \in H^1(\Omega)$ is in $H^{1/2}(\Gamma)$. Thus, there exists a trace lifting $u_g \in H^1(\Omega)$ such that $u_g|_\Gamma = g$. With $u := u_0 + u_g$ we want to find $u_0 \in H_0^1(\Omega)$ such that

$$a(u_0, v) = l(v) - a(u_g, v), \quad \forall v \in H_0^1(\Omega).$$

Proposition 2.3 is still valid, since the bilinear form is bounded.

2.2.2 Neumann Boundary Condition

For the Neumann boundary condition, $\langle \mathbf{A}\nabla u, \mathbf{n}\rangle = g$ on Γ for $g \in L^2(\Gamma)$, the bilinear form is the same as (2.7). The linear form is different, since the boundary integral $\int_\Gamma \langle \mathbf{A}\nabla u, \mathbf{n}\rangle\, v\,\mathrm{d}s$ does not vanish:

$$l(v) := \int_\Omega fv + \int_\Gamma gv. \tag{2.9}$$

The variational formulation in this case states: Find $u \in H^1(\Omega)$ such that

$$\int_\Omega \left(\langle \mathbf{A}\nabla u, \nabla v\rangle + \langle \mathbf{b}, \nabla u\rangle\, v + cuv\right) = \int_\Omega fv + \int_\Gamma gv, \quad \forall v \in H^1(\Omega). \tag{2.10}$$

The following proposition gives the unique solvability:

Proposition 2.4. *Assume that* (2.4) *is fulfilled,* $\mathbf{A}$ *satisfies* (2.2), $\operatorname{div}\mathbf{b} \in L^\infty(\Omega, \mathbb{R})$ *and* $-\frac{1}{2}\operatorname{div}\mathbf{b} + c \geq \delta_0 > 0$. *Further, assume that* $\langle \mathbf{b}, \mathbf{n}\rangle \geq 0$ *on* Γ. *Then, the Neumann problem with variational formulation* (2.10) *has a unique weak solution.*

Proof. We proceed as before in the Dirichlet case and consider $H := H^1(\Omega)$. The proof of the boundedness of the bilinear form does not change.
Since we assumed $\langle \mathbf{b}, \mathbf{n} \rangle \geq 0$ on Γ, we have the following inequality for $u \in H^1(\Omega)$:

$$\begin{aligned}\int_\Omega \langle \mathbf{b}, u\nabla u \rangle &= \frac{1}{2}\int_\Gamma \langle \mathbf{b}, \mathbf{n} \rangle u^2 \,\mathrm{d}s - \int_\Omega \frac{1}{2}\operatorname{div}(\mathbf{b})\, u^2 \\ &\geq -\int_\Omega \frac{1}{2}\operatorname{div}(\mathbf{b})\, u^2\end{aligned}$$

Then, the H^1-ellipticity of the bilinear form follows:

$$\begin{aligned}a(u,u) &\geq \lambda |u|^2_{H^1(\Omega)} + \int_\Omega \left(c - \frac{1}{2}\operatorname{div}(\mathbf{b})\right) u^2 \\ &\geq \lambda |u|^2_{H^1(\Omega)} + \delta_0 \|u\|^2_{L^2(\Omega)} \\ &\geq \min\{\lambda, \delta_0\} \|u\|^2_{H^1(\Omega)}\end{aligned}$$

Finally, l is a linear form $l : H^1(\Omega) \to \mathbb{R}$:

$$\begin{aligned}|l(v)| &= \left| \int_\Omega fv + \int_\Gamma gv \right| \\ &\leq \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} + \|g\|_{L^2(\Gamma)} \|v\|_{L^2(\Gamma)} \\ &\leq \left(\|f\|_{L^2(\Omega)} + C_{T\Gamma} \|g\|_{L^2(\Gamma)} \right) \|v\|_{H^1(\Omega)},\end{aligned}$$

where we used the trace inequality (see Section A.3).
Thus, we can apply the Lax-Milgram Theorem 2.2, which gives us existence and uniqueness of the solution. □

We further want to consider the case $c = 0$ and $\mathbf{b} = \mathbf{0}$; in this setting we have to be a bit more careful. The bilinear form states

$$a(u,v) := \int_\Omega \langle \mathbf{A}\nabla u, \nabla v \rangle \tag{2.11}$$

and we still have the linear form (2.9). The variational formulation is: find $u \in H^1(\Omega)$ such that

$$\int_\Omega \langle \mathbf{A}\nabla u, \nabla v \rangle = \int_\Omega fv + \int_\Gamma gv, \quad \forall v \in H^1(\Omega). \tag{2.12}$$

In this case, the bilinear form is no longer H^1-elliptic, therefore the Lax-Milgram Theorem is not applicable for $H = H^1(\Omega)$. Moreover, the solution of the variational problem (2.12) is only unique up to an additive constant. The appropriate space therefore is the following quotient space (see [13]):

Definition 2.5. *The quotient space*

$$W(\Omega) = H^1(\Omega)/\mathbb{R}$$

is defined as the space of equivalence classes with respect to the relation

$$u \simeq v \iff u - v \text{ is a constant} \quad \forall u, v \in H^1(\Omega).$$

We denote by $\dot{u}$ the class of equivalence represented by u.

Proposition 2.6. *Suppose that Ω is a domain. The following quantity*

$$\|\dot{u}\|_{W(\Omega)} := \|\nabla u\|_{L^2(\Omega)}, \quad \forall u \in \dot{u}, \dot{u} \in W(\Omega),$$

defines a norm on $W(\Omega)$ and $W(\Omega)$ becomes a Banach space. Moreover, $W(\Omega)$ is a Hilbert space with the scalar product

$$(v,w)_{W(\Omega)} = (\nabla v, \nabla w)_{L^2(\Omega)}, \quad \forall v, w \in W(\Omega).$$

Proof. From

$$\|\nabla u\|_{L^2(\Omega)} = 0,$$

it follows

$$u = \text{constant}, \quad \text{i.e. } \dot{u} \simeq 0,$$

meaning that $u \in \dot{0}$, which is sufficient for the proof. The completeness of $W(\Omega)$ is inherited by $H^1(\Omega)$. □

To get existence and uniqueness for the variational formulation: find $\dot{u} \in W(\Omega)$ such that

$$\int_\Omega \langle \mathbf{A}\nabla u, \nabla v\rangle = \int_\Omega fv + \int_\Gamma gv, \quad \forall v \in \dot{v}, \dot{v} \in W(\Omega) \text{ and } \forall u \in \dot{u}, \tag{2.13}$$

we have to assume a **compatibility condition**

$$\int_\Omega f \,\mathrm{d}\mathbf{x} + \int_\Gamma g \,\mathrm{d}s = 0, \tag{2.14}$$

which will become clear in the proof of the next proposition.
Further, we can choose a representative element of the class of equivalence of $\dot{u} \in W(\Omega)$ by fixing the constant mentioned before (see [13]). In particular, we can identify $W(\Omega)$ with the space

$$V(\Omega) := \left\{ v \in H^1(\Omega) \middle| \int_\Omega v \,\mathrm{d}\mathbf{x} = 0 \right\}. \tag{2.15}$$

Therefore, the variational formulation in this case states: Find $u \in V(\Omega)$ such that

$$\int_\Omega \langle \mathbf{A}\nabla u, \nabla v\rangle = \int_\Omega fv + \int_\Gamma gv, \quad \forall v \in V(\Omega), \tag{2.16}$$

where the compatibility condition (2.14) has to be satisfied.

Proposition 2.7. *Assume that* (2.4) *is fulfilled,* **A** *satisfies* (2.2) *and* $g \in L^2(\Gamma)$. *Further, assume that the compatibility condition* (2.14) *is fulfilled. Then, the Neumann problem with variational formulation* (2.16) *has a unique weak solution* $u \in V(\Omega)$.

Proof. Consider $H := V(\Omega)$. Since the space $W(\Omega)$ can be identified with $V(\Omega)$, we verify the conditions of the bilinear form with respect to $W(\Omega)$. The boundedness and the $W(\Omega)$-ellipticity of the bilinear form can be verified immediately:

$$|a(\dot{u}, \dot{v})| \le \Lambda \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} = \Lambda \|\dot{u}\|_{W(\Omega)} \|\dot{v}\|_{W(\Omega)}, \quad \forall u \in \dot{u}, \forall v \in \dot{v},$$

$$a(\dot{u}, \dot{v}) \ge \lambda \|\nabla u\|^2_{L^2(\Omega)} = \lambda \|\dot{u}\|^2_{W(\Omega)}, \quad \forall u \in \dot{u}.$$

For the linear form we have to check that it is well defined on $W(\Omega)$. Consider $v, w \in \dot{v}$ for $\dot{v} \in W(\Omega)$, i.e., $v \simeq w$ and thus $v - w = C$ for some constant C. Then, we have by linearity

$$l(v) - l(w) = \int_\Omega f(v - w) + \int_\Gamma g(v - w) = C\left(\int_\Omega f + \int_\Gamma g \right).$$

With the compatibility condition it follows:

$$l(v) = l(w) \iff v \simeq w,$$

i.e., the linear form is well defined on $W(\Omega)$.
We proceed by considering $V(\Omega)$, therefore we have the Poincaré inequality

$$\|v\|_{L^2(\Omega)} \le C_{P\Omega} \|\nabla v\|_{L^2(\Omega)},$$

for $v \in V(\Omega)$, see also Section A.3. Hence, it follows

$$|l(v)| \le \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} + \|g\|_{L^2(\Gamma)} \|v\|_{L^2(\Gamma)} \le \left(C_{P\Omega} \|f\|_{L^2(\Omega)} + C_{T\Gamma} \|g\|_{L^2(\Gamma)} \right) \|\nabla v\|_{L^2(\Omega)}$$

and we can again apply the Lax-Milgram Theorem 2.2, which gives us existence and uniqueness of the solution. □

In order to solve the Neumann problem for $c = 0$ numerically, we have to perform some further manipulations on the variational formulation (2.16). Consider the idea of [33], where an equivalent saddle point formulation is used, combined with a **Lagrange multiplier** $\lambda \in \mathbb{R}$. We obtain a new variational formulation: Find $u \in H^1(\Omega)$ and $\lambda \in \mathbb{R}$ such that

$$\begin{aligned} a(u,v) + \lambda \int_\Omega v \,\mathrm{d}\mathbf{x} &= l(v) \\ \int_\Omega u \,\mathrm{d}\mathbf{x} &= 0, \end{aligned} \tag{2.17}$$

for all $v \in H^1(\Omega)$. In this formulation the condition that fixes the additive constant is considered as side condition. To get a more suitable version for implementation, we continue as in [33] and apply the test function $v \equiv 1$ to the first equation of (2.17), then we get

$$a(u,1) + \lambda|\Omega| = l(1).$$

Since $c = 0$ and since the compatibility condition holds, it follows $a(u,1) = 0$ and $l(1) = 0$ and therefore $\lambda = 0$. Hence, we can subtract λ from the second equation of (2.17), which leads to

$$\lambda = \int_\Omega u \,\mathrm{d}\mathbf{x}.$$

Finally, we substitute the value of the Lagrange multiplier in the saddle point problem and get the following variational problem: Find $u \in H^1(\Omega)$ such that

$$a(u,v) + \int_\Omega u \,\mathrm{d}\mathbf{x} \int_\Omega v \,\mathrm{d}\mathbf{x} = l(v), \quad \forall v \in H^1(\Omega). \tag{2.18}$$

This is the formulation we consider for numerical treatments and the following proposition shows that it is equivalent to the variational problem (2.16).

Proposition 2.8. *Assume that* (2.4) *is fulfilled,* $\mathbf{A}$ *satisfies* (2.2) *and* $g \in L^2(\Gamma)$*. Then, the Neumann problem with variational formulation* (2.18) *has a unique weak solution* $u \in H^1(\Omega)$*.*
Further, assume that the compatibility condition (2.14) *is satisfied. Then, the variational problem* (2.16) *is equivalent to the modified variational problem* (2.18)*.*

Proof. Consider $H = H^1(\Omega)$. The modified bilinear form is bounded:

$$\begin{aligned} |a(u,v)| &= \left| \int_\Omega \langle \mathbf{A}\nabla u, \nabla v\rangle + \int_\Omega u \int_\Omega v \right| \\ &\le \Lambda \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} + \left| \int_\Omega u \right| \left| \int_\Omega v \right| \\ &\le \Lambda \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} + |\Omega| \, \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \\ &\le \max\{\Lambda, |\Omega|\} \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)}. \end{aligned}$$

The modified bilinear form is $H^1(\Omega)$-elliptic:

$$\begin{aligned} a(u,u) &\ge \lambda \|\nabla u\|^2_{L^2(\Omega)} + \left(\int_\Omega u \right)^2 \\ &\ge \min\{\lambda, 1\} \left(\|\nabla u\|^2_{L^2(\Omega)} + \left(\int_\Omega u \right)^2 \right) \\ &\ge \min\{\lambda, 1\} \frac{1}{1 + C^2_{P\Omega}} \|u\|_{H^1(\Omega)}, \end{aligned}$$

where we used the Poincaré inequality (see Section A.3). Since the linear form did not change, we have again existence and uniqueness.
Consider the solution $u \in H^1(\Omega)$ of problem (2.18), then the equation

$$a(u,v) + \int_\Omega u \,\mathrm{d}\mathbf{x} \int_\Omega v \,\mathrm{d}\mathbf{x} = l(v)$$

holds for all $v \in H^1(\Omega)$. If it holds $u \in V(\Omega)$, then it follows immediately that u solves:

$$a(u,v) = l(v), \quad \forall v \in V(\Omega),$$

which is the variational formulation (2.16).
We can see that u actually lies in $V(\Omega)$ by appending the constant value $v \equiv 1$ to (2.18):

$$|\Omega| \int_\Omega u \,\mathrm{d}\mathbf{x} = \int_\Omega f \,\mathrm{d}\mathbf{x} + \int_\Gamma g \,\mathrm{d}s = 0,$$

since the compatibility condition was assumed.
Further, the other direction is obvious from the derivation of (2.18). □

We can also consider mixed Dirichlet and Neumann boundary conditions. In this case, assume $\Gamma = \Gamma_D \cup \Gamma_N$ with $|\Gamma_D| > 0$ and $\Gamma_D \cap \Gamma_N = \varnothing$. Then, $u = g_D$ on Γ_D, with $g_D \in H^1(\Omega)$ and $\langle \mathbf{A}\nabla u, \mathbf{n}\rangle = g_N$ on Γ_N for $g_N \in L^2(\Gamma_N)$. The propositions for the Neumann boundary condition are still valid, where the assumptions are restricted to Γ_N.

2.2.3 Periodic Boundary Condition

Concerning homogenization problems, periodic functions and periodic coefficients arise. Hence, we have to study so called cell problems and the corresponding Sobolev space.
By $\widehat{\Pi}$ we denote a **reference cell** in $\mathbb{R}^d$, i.e.,

$$\widehat{\Pi} = (0, l_1) \times \cdots \times (0, l_d),$$

where $l_1, \ldots, l_d$ are positive numbers. A function f, defined a.e. on $\mathbb{R}^d$, is called $\widehat{\Pi}$-**periodic** if and only if

$$f(\mathbf{x} + k l_i \mathbf{e}_i) = f(\mathbf{x}) \quad \text{a.e. on } \mathbb{R}^d, \tag{2.19}$$

for any $k \in \mathbb{Z}$ and $i \in \{1, \ldots, d\}$, where $\{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$ is the canonical basis of $\mathbb{R}^d$.
We denote by $H^1_{\mathrm{per}}(\widehat{\Pi})$ the closure of $C^\infty_{\mathrm{per}}(\widehat{\Pi})$ for the H^1-norm, where $C^\infty_{\mathrm{per}}(\widehat{\Pi})$ is the subset of $C^\infty(\mathbb{R}^d)$ of $\widehat{\Pi}$-periodic functions.

Definition 2.9 (Mean value). *The mean value of a function $\phi \in L^1(\Omega)$ over Ω is defined by*

$$\langle \phi \rangle_\Omega = \frac{1}{|\Omega|} \int_\Omega \phi(x) \,\mathrm{d}x.$$

For functions in $H^1(\Omega)$ and with the notion of the mean value, one can prove the Poincaré-Wirtinger inequality (see Theorem A.30) similar to the Poincaré inequality.
For further details about periodic function spaces and the following propositions we refer to [13, Chapters 2,3,4].

Proposition 2.10. *Let $u \in H^1_{\mathrm{per}}(\widehat{\Pi})$. Then, u has the same trace on the opposite faces of $\widehat{\Pi}$.*

The proof uses arguments from the proof of the Trace Theorem and the definition of H^1_{per}, see [13]. In the following, we assume

$$f \in L^2(\widehat{\Pi}), \quad \mathbf{A} = (a_{i,j})_{i,j=1}^d \in L^\infty(\widehat{\Pi}, \mathbb{R}^{d\times d}_{\mathrm{sym}}) \tag{2.20}$$

and $a_{i,j}$ are $\widehat{\Pi}$-periodic for $i, j = 1, \ldots, d$. We consider the problem

$$\begin{aligned} -\operatorname{div}(\mathbf{A}\nabla u) &= f \quad \text{in } \widehat{\Pi}, \\ u &\quad \widehat{\Pi}\text{-periodic.} \end{aligned} \tag{2.21}$$

Hence, we consider the bilinear form (2.11) and the linear form (2.8). A natural space for this problem is $W_{\mathrm{per}}(\widehat{\Pi})$, defined in the spirit of Definition 2.5:

Definition 2.11. *The quotient space*

$$W_{\mathrm{per}}\left(\widehat{\Pi}\right) = H^1_{\mathrm{per}}\left(\widehat{\Pi}\right)/\mathbb{R}$$

is defined as the space of equivalence classes with respect to the relation

$$u \simeq v \iff u - v \text{ is a constant} \quad \forall u, v \in H^1_{\mathrm{per}}\left(\widehat{\Pi}\right).$$

We denote by $\dot{u}$ the class of equivalence represented by u.

Proposition 2.12. *The following quantity*

$$\|\dot{u}\|_{W_{\mathrm{per}}(\widehat{\Pi})} := \|\nabla u\|_{L^2(\widehat{\Pi})}, \quad \forall u \in \dot{u}, \dot{u} \in W_{\mathrm{per}}\left(\widehat{\Pi}\right),$$

defines a norm on $W_{\mathrm{per}}\left(\widehat{\Pi}\right)$. Moreover, the dual space $\left(W_{\mathrm{per}}\left(\widehat{\Pi}\right)\right)'$ can be identified with the set

$$\left\{ l \in \left(H^1_{\mathrm{per}}(\widehat{\Pi})\right)' \mid l(c) = 0, \quad \forall c \in \mathbb{R} \right\},$$

with

$$l(\dot{u}) = l(u) \quad \forall u \in \dot{u}, \dot{u} \in W_{\mathrm{per}}\left(\widehat{\Pi}\right).$$

The proof follows from Proposition 2.6.
For a given f in $\left(W_{\mathrm{per}}\left(\widehat{\Pi}\right)\right)'$, we can again state the variational formulation: Find $\dot{u} \in W_{\mathrm{per}}\left(\widehat{\Pi}\right)$ such that

$$\int_{\widehat{\Pi}} \langle \mathbf{A}\nabla u, \nabla v\rangle = \int_{\widehat{\Pi}} fv, \quad \forall v \in \dot{v}, \dot{v} \in W_{\mathrm{per}}\left(\widehat{\Pi}\right) \text{ and } \forall u \in \dot{u}. \tag{2.22}$$

Similar to the Neumann case, elements of $W_{\mathrm{per}}\left(\widehat{\Pi}\right)$ are defined up to an additive constant, if we fix this constant we can choose a representative element of $\dot{u}$. In particular, we can ask for the solution to have zero mean value. Thus, for f in $\left(W_{\mathrm{per}}\left(\hat{\Pi}\right)\right)'$, we solve the problem

$$\begin{aligned} -\operatorname{div}(\mathbf{A}\nabla u) &= f \quad \text{in } \widehat{\Pi}, \\ u &\quad \widehat{\Pi}\text{-periodic}, \\ \langle u\rangle_{\widehat{\Pi}} &= 0. \end{aligned} \tag{2.23}$$

The variational formulation now takes the form: Find $u \in W_{\mathrm{per}}\left(\widehat{\Pi}\right)$ such that

$$\int_{\widehat{\Pi}} \langle \mathbf{A}\nabla u, \nabla v\rangle = \int_{\widehat{\Pi}} fv, \quad \forall v \in W_{\mathrm{per}}\left(\widehat{\Pi}\right), \tag{2.24}$$

where

$$W_{\mathrm{per}}\left(\widehat{\Pi}\right) = \left\{ v \mid v \in H^1_{\mathrm{per}}\left(\widehat{\Pi}\right), \langle v\rangle_{\widehat{\Pi}} = 0 \right\}. \tag{2.25}$$

Proposition 2.13. *Assume that* (2.20) *is fulfilled,* $\mathbf{A}$ *satisfies* (2.2) *with $\widehat{\Pi}$-periodic coefficients and $f \in \left(W_{\mathrm{per}}\left(\widehat{\Pi}\right)\right)'$. Then, the periodic problem with variational formulation* (2.24) *has a unique weak solution $u \in W_{\mathrm{per}}\left(\widehat{\Pi}\right)$.*

Proof. Consider $H := W_{\mathrm{per}}\left(\widehat{\Pi}\right)$. The boundedness and the $W_{\mathrm{per}}\left(\widehat{\Pi}\right)$-ellipticity of the bilinear form follow in the same manner as in the proof of Proposition 2.7. Further, by Proposition 2.12, the linear form is well defined.
For $v \in W_{\mathrm{per}}\left(\widehat{\Pi}\right)$, since $\langle v\rangle_{\widehat{\Pi}} = 0$, we have the Poincaré-Wirtinger inequality (see Section A.3)

$$\|v\|_{L^2(\widehat{\Pi})} \le C_{PW}\|\nabla v\|_{L^2(\widehat{\Pi})}.$$

Hence, it follows

$$|l(v)| \le \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \le C_{PW}\|f\|_{L^2(\Omega)}\|\nabla v\|_{L^2(\Omega)} = C_{PW}\|f\|_{L^2(\Omega)}\|\dot{v}\|_{W_{\mathrm{per}}(\widehat{\Pi})}$$

and we can again apply the Lax-Milgram Theorem 2.2, which gives us existence and uniqueness of the solution. □

In order to solve the periodic problem numerically, we have to make similar considerations as in the Neumann case. The variational formulation with a Lagrange multiplier $\lambda \in \mathbb{R}$ states: Find $u \in H^1_{\text{per}}\left(\widehat{\Pi}\right)$ and $\lambda \in \mathbb{R}$ such that

$$\begin{aligned} a(u,v) + \lambda \langle v \rangle_{\widehat{\Pi}} &= l(v) \\ \langle u \rangle_{\widehat{\Pi}} &= 0, \end{aligned} \tag{2.26}$$

for all $v \in H^1_{\text{per}}$. Again, we apply the test function $v \equiv 1$ to the first equation of (2.26), then we get

$$a(u,1) + \lambda = l(1).$$

Due to the definition of the bilinear form and since $f \in \left(W_{\text{per}}\left(\widehat{\Pi}\right)\right)'$, it follows $a(u,1) = 0$ and $l(1) = 0$ and therefore $\lambda = 0$. Hence, we can subtract λ from the second equation of (2.26), which leads to

$$\lambda = \langle u \rangle_{\widehat{\Pi}}.$$

Finally, we substitute the value of the Lagrange multiplier in the saddle point problem and get the following variational problem: Find $u \in H^1_{\text{per}}\left(\widehat{\Pi}\right)$ such that

$$a(u,v) + \langle u \rangle_{\widehat{\Pi}} \langle v \rangle_{\widehat{\Pi}} = l(v), \quad \forall v \in H^1_{\text{per}}\left(\widehat{\Pi}\right). \tag{2.27}$$

This is the formulation we consider for numerical treatments and the following proposition shows that it is equivalent to the variational problem (2.24).

Proposition 2.14. *Assume that* (2.20) *is fulfilled,* $\mathbf{A}$ *is* $\widehat{\Pi}$*-periodic and satisfies* (2.2)*, and* $f \in \left(H^1_{\text{per}}\left(\widehat{\Pi}\right)\right)'$*. Then, the periodic problem with variational formulation* (2.27) *has a unique weak solution* $u \in H^1_{\text{per}}\left(\widehat{\Pi}\right)$*.*

Further, assume that $f \in \left(W_{\text{per}}\left(\widehat{\Pi}\right)\right)'$*. Then, the variational problem* (2.24) *is equivalent to the modified variational problem* (2.27)*.*

Proof. Consider $H = H^1_{\text{per}}\left(\widehat{\Pi}\right)$. The modified bilinear form is bounded:

$$\begin{aligned} |a(u,v)| &\leq \Lambda \|\nabla u\|_{L^2(\widehat{\Pi})} \|\nabla v\|_{L^2(\widehat{\Pi})} + \frac{1}{\left|\widehat{\Pi}\right|} \|u\|_{L^2(\widehat{\Pi})} \|v\|_{L^2(\widehat{\Pi})} \\ &\leq \max\left\{\Lambda, \frac{1}{\left|\widehat{\Pi}\right|}\right\} \|u\|_{H^1(\widehat{\Pi})} \|v\|_{H^1(\widehat{\Pi})}. \end{aligned}$$

The modified bilinear form is $H^1_{\text{per}}\left(\widehat{\Pi}\right)$-elliptic:

$$\begin{aligned} a(u,u) &\geq \min\left\{\lambda, \frac{1}{|\widehat{\Pi}|^2}\right\} \left(\|\nabla u\|^2_{L^2(\Omega)} + \left(\int_\Omega u \right)^2 \right) \\ &\geq \min\left\{\lambda, \frac{1}{|\widehat{\Pi}|^2}\right\} \frac{1}{1 + C^2_{P\Omega}} \|u\|_{H^1(\Omega)}, \end{aligned}$$

where we used the Poincaré inequality (see Section A.3). Since the linear form did not change, we have again existence and uniqueness.
Consider the solution $u \in H^1_{\text{per}}\left(\widehat{\Pi}\right)$ of problem (2.27), then the equation

$$a(u,v) + \langle u \rangle_{\widehat{\Pi}} \langle v \rangle_{\widehat{\Pi}} = l(v)$$

holds for all $v \in H^1_{\text{per}}\left(\widehat{\Pi}\right)$. If it holds $u \in W_{\text{per}}\left(\widehat{\Pi}\right)$, then it follows immediately that u solves:

$$a(u,v) = l(v), \quad \forall v \in W_{\text{per}}\left(\widehat{\Pi}\right),$$

which is the variational formulation (2.24).
We can see that u actually lies in $W_{\text{per}}\left(\widehat{\Pi}\right)$ by appending the constant value $v \equiv 1$ to (2.27):

$$\langle u \rangle_{\widehat{\Pi}} = l(1) = 0,$$

since $f \in \left(W_{\text{per}}\left(\widehat{\Pi}\right)\right)'$.
Further, the other direction is obvious from the derivation of (2.27). □

The cell problems defined in the next chapter only contain the principal part with coefficient $\mathbf{A}$, therefore we considered $\mathbf{b} = \mathbf{0}$ and $c = 0$ in this section. One could prove existence and uniqueness of periodic problems with terms $\mathbf{b}$ and c similar to the Neumann case. Further, the right-hand side of the cell problems will take a special form, in particular f will be of the form $-\operatorname{div}(\mathbf{h})$ for $\mathbf{h} \in L^2(\widehat{\Pi}, \mathbb{R}^d)$. Consider the following relation:

$$\int_{\widehat{\Pi}} -\operatorname{div}(\mathbf{h})\, v = \int_{\widehat{\Pi}} \langle \mathbf{h}, \nabla v \rangle ,$$

which holds, since

$$\int_{\partial\widehat{\Pi}} \langle \mathbf{h}, \mathbf{n} \rangle\, v = 0, \quad \forall v \in W_{\mathrm{per}}\left(\widehat{\Pi}\right),$$

due to Proposition 2.10. Then, the linear form is

$$l(v) := \int_{\widehat{\Pi}} \langle \mathbf{h}, \nabla v \rangle .$$

2.3 Finite Element Method

Before we can treat elliptic boundary value problems numerically, we first have to explain the concept of the Galerkin discretization and the finite elements. Further, we state important a priori error estimates.

2.3.1 Galerkin Method

Let $\Omega \subset \mathbb{R}^d$ be an open, bounded domain with Lipschitz boundary Γ. We consider the variational problem:

$$\text{Find } u \in V \text{ such that } a(u, v) = l(v), \quad \forall\, v \in V, \tag{2.28}$$

for a bounded, V-elliptic bilinear form $a : V \times V \to \mathbb{R}$, a linear form $l : V \to \mathbb{R}$ and a suitable space $V \subset H^1(\Omega)$. The coefficients satisfy the conditions $f \in L^2(\Omega)$, $\mathbf{A} \in L^\infty\left(\Omega, \mathbb{R}^{d\times d}_{\mathrm{sym}}\right)$, $\mathbf{b} \in L^\infty\left(\Omega, \mathbb{R}^d\right)$ and $c \in L^\infty(\Omega, \mathbb{R}_{\geq 0})$. Further, $\mathbf{A}$ is assumed to be uniformly elliptic and $\mathbf{b}$ and c fulfil additional conditions depending on the problem considered, as explained in the section before.
For the numerical approximation of the solution we consider the variational problem on some suitable finite-dimensional subspace $V_N \subset V$, with $\dim V_N = N < \infty$. The Galerkin solution $u^N \in V_N$ then satisfies

$$a\left(u^N, v^N\right) = l\left(v^N\right), \quad \forall v^N \in V_N. \tag{2.29}$$

Suppose $\{\psi_1, \psi_2, \ldots, \psi_N\}$ is a basis for V_N and assume that u^N has the form

$$u^N = \sum_{j=1}^{N} u_j \psi_j .$$

Then, (2.29) is equivalent to the system of equations:

$$\sum_{j=1}^{N} a(\psi_j, \psi_i)\, u_j = l(\psi_i), \quad i = 1, 2, \ldots, N,$$

which can be written as

$$\mathbf{L}\mathbf{u} = \mathbf{f}, \tag{2.30}$$

with the system matrix

$$\mathbf{L} = (l_{i,j})_{i,j=1}^{N} \in \mathbb{R}^{N\times N}, \qquad l_{i,j} = a(\psi_j, \psi_i), \quad \forall\, 1 \leq i, j \leq N,$$

and the right-hand side

$$\mathbf{f} = (f_i)_{i=1}^{N} \in \mathbb{R}^N, \qquad f_i = l(\psi_i), \quad \forall\, 1 \leq i \leq N,$$

where $\mathbf{u} = (u_i)_{i=1}^{N}$.

Remark 2.15. *Since a is a V-elliptic bilinear form, the matrix $\mathbf{L}$ is positive definite*[1]*. Furthermore, if a is a symmetric bilinear form, then the matrix $\mathbf{L}$ is symmetric.*

Remark 2.16. *The existence and uniqueness of the solution u^N of (2.29) is given by the Lax-Milgram Theorem 2.2. Further, the solution u^N is stable, i.e., satisfies*

$$\left\|u^N\right\|_V \leq \frac{1}{\alpha^{\text{ell}}} \|l\|_{V'}.$$

The following statement will be important for establishing error bounds for finite element approximations:

Theorem 2.17 (Céa's Lemma). *Suppose the bilinear form a is bounded and V-elliptic. In addition, suppose $u \in V$ and $u^N \in V_N \subset V$ are the solutions of the variational problem (2.28) and (2.29), respectively. Then,*

$$\left\|u - u^N\right\|_V \leq \frac{\alpha^{\text{cont}}}{\alpha^{\text{ell}}} \inf_{v^N \in V_N} \left\|u - v^N\right\|_V .$$

A proof can be found in [9, Chapter II.4].

Remark 2.18. *Let u and u^N be the solutions of the variational problem in V and $V_N \subset V$, respectively. Then it holds*

$$a\left(u - u^N, v\right) = 0, \quad \forall v \in V_N,$$

*which is called **Galerkin orthogonality**.*

From Céa's Lemma it follows, that it is essential to choose the function space V_N well in order to get an accurate approximation of the numerical solution. We will see in the next subsections, that it is possible to take piecewise polynomials and the desired accuracy is then achieved by a sufficiently fine partition of Ω.

2.3.2 Finite Elements

As explained before, we will solve the variational problem on some space V_N, which is called finite element space in practice. According to Céa's Lemma the accuracy of the numerical solution depends on the choice of this trial space. For simplicity, we explain the construction of the finite element space only in two dimensions and we consider a convex, polygonal domain Ω. Higher dimensions are well discussed in the literature.
We consider a partition of the domain into a finite number of triangles, also called elements, which we always assume to be closed. Then, we can choose trial functions, defined on each element by a polynomial of given degree. For that, the set of polynomials of degree $\leq t$ is denoted by

$$\mathcal{P}_t := \left\{ u(x_1, x_2) = \sum_{\substack{i+k \leq t \\ i,k \geq 0}} c_{i,k} x_1^i x_2^k \right\}. \tag{2.31}$$

Further, note that we consider conforming elements, i.e., functions that lie in the Sobolev space in which the variational formulation is stated.

Definition 2.19 (Admissible triangulation). *A partition $\mathcal{T} = \{T_1, T_2, \ldots, T_M\}$ of Ω into closed triangular elements is called admissible, if the following properties hold:*

a) $\overline{\Omega} = \bigcup_{i=1}^{M} T_i$.

b) If $T_i \cap T_j$ consists of exactly one point, then it is a common vertex of T_i and T_j.

c) If, for $i \neq j$, $T_i \cap T_j$ consists of more than one point, then $T_i \cap T_j$ is a common edge of T_i and T_j.

[1] A matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ is called positive definite, if it holds $\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle > 0$ for all $\mathbf{x} \in \mathbb{R}^d \backslash \{\mathbf{0}\}$.

Definition 2.20. *For every triangle $T \in \mathcal{T}$, we define by h_T the length of the largest edge (diameter) and by ρ_T the diameter of the largest inscribed circle (see Figure 2.1). We write $\mathcal{T}_h$ instead of $\mathcal{T}$ if every element has diameter at most $2h$, where $h = \max_{T \in \mathcal{T}} h_T$.*

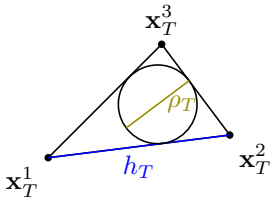


Figure 2.1: Example of a triangle T with h_T and ρ_T.

Definition 2.21 (Shape regular, uniform). *The triangulation $\mathcal{T}_h$ is called*

a) shape regular, if there exists a constant $C > 0$ such that

$$\rho_T \geq h_T / C.$$

b) uniform, if there exists a constant $C > 0$ such that

$$\rho_T \geq h / C.$$

The uniformity is a stronger requirement than the shape regularity, for an example see Figure 2.2, where for the triangulation on the left, both constants are approximately 1.3. For the triangulation on the right, the shape regularity constant is again 1.3, but the uniformity constant gets very large.

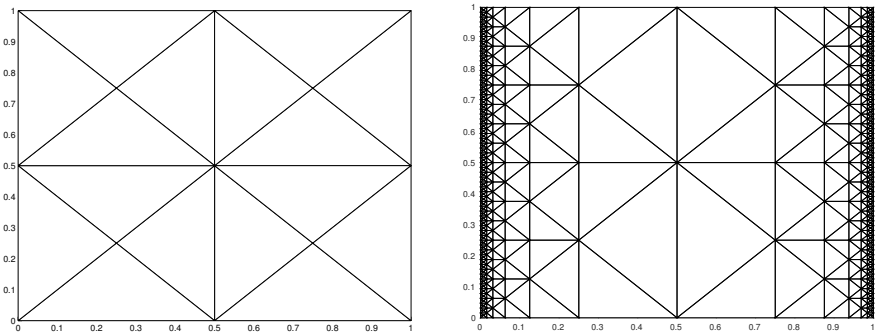


Figure 2.2: A uniform and a shape regular, non-uniform triangulation for $h = 0.5$.

The following theorem shows that if we choose piecewise polynomials as trial functions, they further have to be globally continuous in order to be in $H^1(\Omega)$.

Theorem 2.22. *Let $k \geq 1$ and suppose Ω is bounded. Then, a function $v : \overline{\Omega} \to \mathbb{R}$, such that $v|_T \in C^k(T)$, $\forall T \in \mathcal{T}_h$, belongs to $H^k(\Omega)$ if and only if $v \in C^{k-1}\left(\overline{\Omega}\right)$.*

A proof can be found in [9, Chapter II.5].
Since $V_N \subset H^1(\Omega)$, continuous elements are suitable in our case and we can write:

$$V_h := V_N := \left\{\psi \in C^0\left(\overline{\Omega}\right) \mid \forall T \in \mathcal{T}_h : \psi|_T \in \mathcal{P}_t\right\} \cap V. \tag{2.32}$$

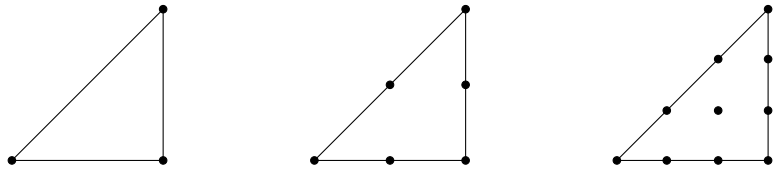
As mentioned before, we restrict ourselves to triangular elements. Further, note that the set of polynomials $\mathcal{P}_t$ is invariant under affine linear transformations.

Remark 2.23. *Let $t \geq 0$. Given a triangle T, suppose $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_s$ are the $s = \binom{t+2}{2}$ points in T which lie on $t+1$ equidistant lines parallel to one edge of T, independent of the choice of the edge. For a triangle as in Figure 2.3, these s nodes are defined by the set*

$$\Theta_t := \left\{ \frac{\boldsymbol{\alpha}}{t} \,\middle|\, \boldsymbol{\alpha} = (\alpha_1, \alpha_2) \in \mathbb{N}_0^2 \text{ with } \alpha_2 - \alpha_1 \leq 0 \text{ and } \alpha_1, \alpha_2 \leq t \right\}.$$

Then, for every $v \in C^0(T)$, there is a unique polynomial p of degree $\leq t$ satisfying the interpolation conditions

$$p(\mathbf{x}_i) = v(\mathbf{x}_i), \quad i = 1, 2, \ldots, s.$$



$t = 1$ $t = 2$ $t = 3$

Figure 2.3: Local degree of freedoms for the linear, quadratic and cubic Lagrange basis.

The construction of the C^0-elements works as follows, for further details see [9, Chapter II.5]. Let $t \geq 1$ and consider a triangulation $\mathcal{T}_h$ of Ω. We place $s = \binom{t+2}{2}$ points in each triangle T, so that they belong to a set Θ_t corresponding to T. With Remark 2.23, we have a unique polynomial on each triangle, by choosing values at these points. The restriction of any such polynomial to an edge is a polynomial of degree $\leq t$ in one variable. Since the two polynomials of the same order corresponding to the triangles containing this edge interpolate the same value at the $t+1$ points on this edge, they must reduce to the same one-dimensional polynomial. This ensures the global continuity of our elements.

Definition 2.24 (Lagrange basis). *Polynomials of degree $\leq t$, which take the value one at exactly one of the $s = \binom{t+2}{2}$ points and vanish at all the others, form a nodal basis for V_N, also called Lagrange basis.*

Now, we state the formal construction of a finite element and the finite element space in $\mathbb{R}^2$, for $\mathbb{R}^d$ see [9, Chapter II.5].

Definition 2.25 (Finite element space). *A family of finite element spaces V_h, for partitions $\mathcal{T}_h$ of $\Omega \subset \mathbb{R}^2$, is called an affine family, provided there exists a finite element $(T_{\text{ref}}, \mathcal{P}_{\text{ref}}, \Sigma)$, called the reference element, with the following properties:*

a) T_{ref} is a closed polygon in $\mathbb{R}^2$.

b) $\mathcal{P}_{\text{ref}}$ is a subspace of $C^0(T)$ with finite dimension s.

c) Σ is a set of s linearly independent functionals on $\mathcal{P}_{\text{ref}}$. Every $p \in \mathcal{P}_{\text{ref}}$ is uniquely defined by the value of the s functionals in Σ. (Interpolation condition)

d) For every $T \in \mathcal{T}_h$, there exists an affine mapping $\chi_T : T_{\text{ref}} \to T$ such that for every $v \in V_h$, its restriction to T has the form

$$v(\mathbf{x}) = p\left(\chi_T^{-1}(\mathbf{x})\right), \quad \text{with } p \in \mathcal{P}_{\text{ref}}.$$

For linear finite elements, we have $\mathcal{P}_{\text{ref}} = \mathcal{P}_1$ and the local degree of freedom is $s = 3$. We always consider $T_{\text{ref}} = \{(0,0), (1,0), (1,1)\}$ and one can easily check that $\widehat{\psi}_1(\hat{\mathbf{x}}) = 1 - \hat{x}_1$, $\widehat{\psi}_2(\hat{\mathbf{x}}) = \hat{x}_1 - \hat{x}_2$ and $\widehat{\psi}_3(\hat{\mathbf{x}}) = \hat{x}_2$ are nodal basis functions on T_{ref}, hence $\Sigma = \left\{\widehat{\psi}_1, \widehat{\psi}_2, \widehat{\psi}_3\right\}$. Then, a polynomial $p \in \mathcal{P}_1$ can be written as

$$p(\hat{\mathbf{x}}) = \sum_{i=1}^{3} p_i \widehat{\psi}_i(\hat{\mathbf{x}}), \quad \text{with } (p_i)_{i=1}^3 \in \mathbb{R}^3.$$

Note that for every $T \in \mathcal{T}_h$ and every $p \in \mathcal{P}_1$ the polynomial p is uniquely defined by its values in the three vertices of T. Hence, each function of V_h is uniquely determined by its node values.
Each triangle $T \in \mathcal{T}_h$ can be mapped by a (non-unique) affine transformation onto the initial triangle T_{ref}. The affine transformation is given by

$$\begin{aligned} &\chi_T : T_{\text{ref}} \to T \\ &\qquad \hat{\mathbf{x}} \to \mathbf{x} = \chi_T(\hat{\mathbf{x}}) := \mathbf{x}_A + \hat{x}_1(\mathbf{x}_B - \mathbf{x}_A) + \hat{x}_2(\mathbf{x}_C - \mathbf{x}_B) = (\mathrm{D}\chi_T)\,\hat{\mathbf{x}} + \mathbf{x}_A, \end{aligned} \tag{2.33}$$

where $\mathbf{x}_A$, $\mathbf{x}_B$ and $\mathbf{x}_C$ are the vertices of T and $(0,0)$, $(1,0)$ and $(1,1)$ are the vertices of T_{ref}, see also Figure 2.4.

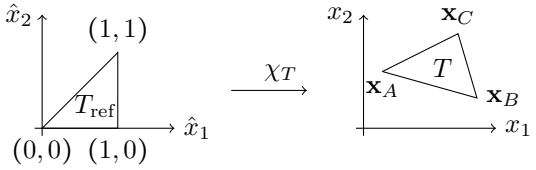


Figure 2.4: Affine mapping χ_T from the reference triangle to an arbitrary triangle.

We can write down the Jacobian matrix explicitly

$$\mathrm{D}\chi_T = \begin{pmatrix} x_B - x_A & x_C - x_B \\ y_B - y_A & y_C - y_B \end{pmatrix} = \begin{pmatrix} \partial_{\hat{x}_1} x_1 & \partial_{\hat{x}_2} x_1 \\ \partial_{\hat{x}_1} x_2 & \partial_{\hat{x}_2} x_2 \end{pmatrix} \tag{2.34}$$

and see directly that the determinant is twice the area of T:

$$\det(\mathrm{D}\chi_T) = (\mathbf{x}_B - \mathbf{x}_A) \times (\mathbf{x}_C - \mathbf{x}_B) = 2|T|. \tag{2.35}$$

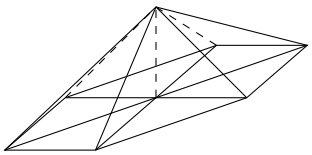
Since the area of the reference triangle is exactly $\frac{1}{2}$, this can be written as:

$$\det(\mathrm{D}\chi_T) = \frac{|T|}{|T_{\text{ref}}|}.$$

Those considerations will be useful for the numerical computation of the system matrix, since the integrals will only be computed on the reference triangle.
Further, one can show the following estimates:

$$\|\mathrm{D}\chi_T\| \le \frac{h_T}{\rho_{T_{\text{ref}}}}, \quad \left\|(\mathrm{D}\chi_T)^{-1}\right\| \le \frac{\rho_T}{h_{T_{\text{ref}}}}.$$

A global basis function ψ corresponding to a global node then has the following support:



2.3.3 A Priori Error Estimates

In this section we give error bounds for finite element approximations, where we consider $H_0^m(\Omega) \subset V \subset H^m(\Omega)$, for $m \in \mathbb{N}$. They are not derived for every element, but for a reference element and then with transformation formulas carried onto general shape regular grids. The error for an interpolation method provides an upper bound for the error of the best approximation. We summarize in the following the main results of [9, Chapter II.6], where all the proofs can be found.

Recall that we have from the Sobolev Embedding Theorem A.21, that H^m is continuously embedded in C^0 for $m \geq 2$. The interpolation operator $I_h : H^m(\Omega) \to V_h$ can be defined as

$$I_h(u) = \sum_{i=1}^{N} u(\mathbf{x}_i)\psi_i, \tag{2.36}$$

where $N = \dim(V_h)$, $m \geq 2$ and $\mathbf{x}_i$ as in Remark 2.23.

Lemma 2.26. *Let $t \geq 2$ and $\Omega \subset \mathbb{R}^2$ be a polygonal domain. Suppose $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_s$ are $s = t(t+1)/2$ prescribed points in $\overline{\Omega}$ such that the interpolation operator $I_h : H^t(\Omega) \to \mathcal{P}_{t-1}$ is well defined. Then there exists a constant c_I, depending on Ω and the nodes $\mathbf{x}_i$, such that*

$$\|u - I_h u\|_{H^t(\Omega)} \leq c_I |u|_{H^t(\Omega)}, \quad \forall u \in H^t(\Omega).$$

Theorem 2.27. *Let $t \geq 2$ and $0 \leq m \leq t$. Suppose $\mathcal{T}_h$ is a shape regular triangulation of Ω. Then, there exists a constant c_I, depending on Ω, t and the shape regularity constant, such that*

$$\|u - I_h u\|_{H^m(\Omega)} \leq c_I h^{t-m} |u|_{H^t(\Omega)} \quad \text{for } u \in H^t(\Omega),$$

where I_h denotes interpolation by a piecewise polynomial of degree $t-1$.

Remark 2.28. *For $u \in H^2(\Omega)$ and continuous, piecewise linear trial functions on triangles, we therefore have the following estimates:*

$$\|u - I_h u\|_{H^1(\Omega)} \leq c_I h |u|_{H^2(\Omega)},$$
$$\|u - I_h u\|_{L^2(\Omega)} \leq c_I h^2 |u|_{H^2(\Omega)}.$$

Theorem 2.29 (Inverse estimates). *Let (V_h) be an affine family of finite elements consisting of piecewise polynomials of degree k associated with uniform partitions. Then there exists a constant c_{inv}, depending on k, t and the uniformity constant, such that for all $0 \leq m \leq t$,*

$$\|v_h\|_{H^t(\Omega)} \leq c_{\text{inv}} h^{m-t} \|v_h\|_{H^m(\Omega)}, \quad \forall v_h \in V_h.$$

Remark 2.30. *For continuous, piecewise linear trial functions on triangles, we therefore have the following estimate:*

$$\|u_h\|_{H^1(\Omega)} \leq c_{\text{inv}} h^{-1} \|u_h\|_{L^2(\Omega)}, \quad \forall u_h \in V_h.$$

The interpolation operator I_h can only be applied to H^2 functions, for H^1 functions the Clément interpolation operator $\mathbf{C}_h$ is applicable.
For a shape regular triangulation $\mathcal{T}_h$ of Ω we consider the following notation, see also Figure 2.5. For a node $\mathbf{x}_i$ we define the **support** of ψ_i by:

$$\omega_i := \omega_{\mathbf{x}_i} := \bigcup \{T \in \mathcal{T}_h \mid \mathbf{x}_i \in T\}. \tag{2.37}$$

The support of a triangle T, also called **patch**, is defined by:

$$\omega_T := \bigcup \{\omega_i \mid \mathbf{x}_i \in T\}. \tag{2.38}$$

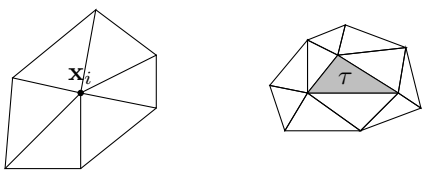


Figure 2.5: Support of a node $\mathbf{x}_i$ and of an element T.

The Clément interpolation operator $\mathbf{C}_h : H^1(\Omega) \to V_h \subset H^1(\Omega)$ is defined by

$$\mathbf{C}_h(u) := \sum_{i=1}^{N} \gamma_i(u)\psi_i,$$

where $\gamma_i : L^2(\omega_i) \to \mathcal{P}_0$ is a local L^2-projection onto the constant functions. The operator γ_i has to be modified in the case of a homogeneous Dirichlet boundary condition, see [9, p. 85]. We have the following local estimate:

Proposition 2.31. *Let $\mathcal{T}_h$ be a shape regular triangulation of Ω. Then there exists a linear mapping $\mathbf{C}_h : H^1(\Omega) \to V_h$ such that*

$$\|v - \mathbf{C}_h v\|_{H^m(T)} \le c h_T^{1-m} \|v\|_{H^1(\omega_T)} \quad \text{for } v \in H^1(\Omega),\ m = 0, 1,\ T \in \mathcal{T}_h.$$

For estimates on the domain Ω, similar to Lemma 2.26, see Section A.4.
By duality technique we can extend the derived estimates from the energy norm to the L^2-norm. The aim is to get a priori bounds of the form

$$\|u - u_h\| \le c h^p, \tag{2.39}$$

where u is the true solution and u_h an approximation in V_h. By p we denote the order of approximation, which depends on the considered Sobolev norm, the degree of the polynomials in the finite elements and the regularity of the solution which we explain in the following.

Definition 2.32 (H^s-regularity). *Let $m \ge 1$, $H_0^m(\Omega) \subset V \subset H^m(\Omega)$, and suppose a is a V-elliptic bilinear form. Then, for $s \ge 2m$, the variational problem*

$$a(u, v) = l(v), \quad \forall v \in V,$$

is called H^s-regular provided that there exists a constant c_R, depending on Ω, a and s, such that for every $f \in H^{s-2m}(\Omega)$, there is a solution $u \in H^s(\Omega)$ with

$$\|u\|_{H^s(\Omega)} \le c_R \|f\|_{H^{s-2m}(\Omega)}. \tag{2.40}$$

The regularity result is only given for a homogeneous Dirichlet boundary condition. More general results, such as mixed Dirichlet and Neumann boundary conditions will not be discussed in this thesis.

Theorem 2.33. *Let a be an H_0^1-elliptic bilinear form with sufficiently smooth coefficient functions. If Ω is convex, then the Dirichlet problem is H^2-regular.*

A proof can be found in [20].

Theorem 2.34. *Suppose $\mathcal{T}_h$ is a family of shape regular triangulations of Ω, where Ω is a convex, polygonal domain. Then, the finite element approximation $u_h \in V_h$, for polynomial degree $t \ge 1$, satisfies*

$$\|u - u_h\|_{H^1(\Omega)} \le \frac{\alpha^{\text{cont}}}{\alpha^{\text{ell}}} c_I h \|u\|_{H^2(\Omega)} \le \frac{\alpha^{\text{cont}}}{\alpha^{\text{ell}}} c_I c_R h \, \|f\|_{L^2(\Omega)}.$$

The proof of the L^2-estimate requires a duality argument, which has been called Aubin-Nitsche's trick:

Lemma 2.35 (Aubin-Nitsche Lemma). *Let H and V be Hilbert spaces, such that $V \hookrightarrow H$. Then the finite element solution in $V_h \subset V$ satisfies*

$$\|u - u_h\|_H \le \alpha^{\text{cont}} \|u - u_h\|_V \sup_{g \in H} \left\{ \frac{1}{\|g\|_H} \inf_{v \in V_h} \|\varphi_g - v\|_V \right\},$$

where for every $g \in H$, $\varphi_g \in V$ denotes the corresponding unique (weak) solution of the equation

$$a(w, \varphi_g) = (g, w)_H, \quad \forall w \in V.$$

Theorem 2.36. *Under the hypothesis of Theorem 2.34, if $u \in H^1(\Omega)$ is the solution of the associated variational problem, then*

$$\|u - u_h\|_{L^2(\Omega)} \leq \alpha^{\text{cont}} c_I c_R h \|u - u_h\|_{H^1(\Omega)}.$$

If in addition $f \in L^2(\Omega)$ so that $u \in H^2(\Omega)$, then

$$\|u - u_h\|_{L^2(\Omega)} \leq \frac{\left(\alpha^{\text{cont}}\right)^2}{\alpha^{\text{ell}}} c_I^2 c_R^2 h^2 \|f\|_{L^2(\Omega)}. \tag{2.41}$$

In conclusion, we get quadratic convergence in the L^2-norm and linear convergence in the H^1-norm, for the homogeneous Dirichlet problem. In the numerical examples we will see, that this also holds for the other boundary value problems.

2.4 A Posteriori Error Estimation

We introduce here the a posteriori error estimates of functional type for a general reaction-convection-diffusion problem. The practical implementation will be discussed in Chapter 5. In Chapter 4 we will derive a posteriori error estimates for homogenization problems, based on the subsequent theory.
Following [26, Section 4.3], we consider a general reaction-convection-diffusion problem

$$\begin{aligned} -\operatorname{div}(\mathbf{A}\nabla u) + \langle \mathbf{b}, \nabla u\rangle + cu &= f \quad && \text{in } \Omega, \\ u &= g_D && \text{on } \Gamma_D, \\ \langle \mathbf{A}\nabla u, \mathbf{n}\rangle &= g_N && \text{on } \Gamma_N, \end{aligned} \tag{2.42}$$

where $\Omega \subset \mathbb{R}^d$ is a bounded domain with Lipschitz continuous boundary Γ, that consists of two measurable non intersecting parts Γ_D and Γ_N. We assume that $|\Gamma_D| > 0$, $g_D \in H^1(\Omega)$ and the matrix $\mathbf{A} \in L^\infty\left(\Omega, \mathbb{R}^{d\times d}\right)$ is symmetric and satisfies the relation

$$\alpha^{\text{ell}} \|\xi\|_2^2 \leq \langle \mathbf{A}(\mathbf{x})\xi, \xi\rangle \leq \alpha^{\text{cont}} \|\xi\|_2^2 \quad \text{for all } \xi \in \mathbb{R}^d \text{ and } \mathbf{x} \in \Omega.$$

We further assume that $\mathbf{b} \in L^\infty\left(\Omega, \mathbb{R}^d\right)$, with $\operatorname{div}\mathbf{b} \in L^\infty(\Omega, \mathbb{R})$, $f \in L^2(\Omega)$, $g_N \in L^2(\Gamma_N)$, $c \in L^\infty(\Omega, \mathbb{R})$ and

$$-\frac{1}{2}\operatorname{div}\mathbf{b} + c =: \delta^2 \geq \delta_0^2.$$

One more assumption is, that the function $\kappa(\mathbf{x}) := \frac{1}{2}\langle \mathbf{b}, \mathbf{n}\rangle(\mathbf{x})$ is defined at almost all points of Γ and the inflow part of the boundary is a subset of Γ_D, i.e.,

$$\Gamma^- := \{\mathbf{x} \in \Gamma \mid \kappa(\mathbf{x}) < 0\} \subset \Gamma_D.$$

Let

$$g_D + V_0 := \{w = u + w_0 \mid w_0 \in V_0(\Omega)\},$$

where

$$V_0 := \left\{w \in H^1(\Omega) \mid w = 0 \text{ on } \Gamma_D\right\}.$$

The generalized solution of (2.42) is a function in $g_D + V_0$, satisfying the integral identity

$$\int_\Omega \{\langle \mathbf{A}\nabla u, \nabla w\rangle + \langle \mathbf{b}, \nabla u\rangle w + cuw\}\, \mathrm{d}\mathbf{x} = \int_\Omega fw\, \mathrm{d}\mathbf{x} + \int_{\Gamma_N} g_N w\, \mathrm{d}s, \quad \forall w \in V_0. \tag{2.43}$$

With the above assumptions, Proposition 2.4 shows that the generalized solution of (2.43) exists and is unique.
In the following theorem we state a general form of a computable upper bound of the error measured in a natural energy type norm.

Theorem 2.37. *Let the assumptions on the reaction-convection-diffusion problem be true. Then, for any $v \in g_D + V_0$ and $\mathbf{y} \in H_{\Gamma_N}(\Omega, \operatorname{div})$ the following estimate holds:*

$$|[u-v]| \le \mathcal{M}(v,\mathbf{y}) := \|\mathbf{y} - \mathbf{A}\nabla v\|_{\mathbf{A}^{-1}} + \frac{1}{\sqrt{\alpha^{\mathrm{ell}}}}\left(C_{F\Omega}\,\|r_\Omega(v,\mathbf{y})\|_{L^2(\Omega)} + C_{T\Gamma_N}\,\|g_N - \langle \mathbf{y},\mathbf{n}\rangle\|_{L^2(\Gamma_N)}\right),$$

where

$$\begin{aligned} r_\Omega(v,\mathbf{y}) &:= f - \langle \mathbf{b}, \nabla v\rangle - cv + \operatorname{div}\mathbf{y}, \\ |[u-v]|^2 &:= \|\nabla(u-v)\|_{\mathbf{A}}^2 + \int_\Omega \delta^2 (u-v)^2\,\mathrm{d}\mathbf{x} + \int_{\Gamma_N} \kappa (u-v)^2\,\mathrm{d}s \end{aligned}$$

and $C_{F\Omega}$ and $C_{T\Gamma_N}$ are constants in the Friedrichs type inequalities, see Section A.3.

Proof. Rewrite the variational formula, by taking $w = u - v$ and subtracting the left-hand side for $u = v$:

$$\begin{aligned} &\int_\Omega \left\{\langle \mathbf{A}\nabla(u-v), \nabla(u-v)\rangle + \langle \mathbf{b}, \nabla(u-v)\rangle (u-v) + c(u-v)^2\right\}\mathrm{d}\mathbf{x} \\ &= \int_\Omega (f - \langle \mathbf{b}, \nabla v\rangle - cv)(u-v)\,\mathrm{d}\mathbf{x} - \int_\Omega \langle \mathbf{A}\nabla v, \nabla(u-v)\rangle\,\mathrm{d}\mathbf{x} + \int_{\Gamma_N} g_N (u-v)\,\mathrm{d}s. \end{aligned} \tag{2.44}$$

With the equation

$$\begin{aligned} \int_\Omega \langle \mathbf{b}, \nabla(u-v)\rangle (u-v)\,\mathrm{d}\mathbf{x} &= \frac{1}{2}\int_\Omega \left\langle \mathbf{b}, \nabla(u-v)^2\right\rangle \mathrm{d}\mathbf{x} \\ &= -\frac{1}{2}\int_\Omega (\operatorname{div}\mathbf{b})(u-v)^2\,\mathrm{d}\mathbf{x} + \frac{1}{2}\int_\Omega \operatorname{div}\left(\mathbf{b}(u-v)^2\right)\mathrm{d}\mathbf{x} \\ &= -\frac{1}{2}\int_\Omega (\operatorname{div}\mathbf{b})(u-v)^2\,\mathrm{d}\mathbf{x} + \frac{1}{2}\int_{\Gamma_N} \langle \mathbf{b},\mathbf{n}\rangle (u-v)^2\,\mathrm{d}s \\ &= -\frac{1}{2}\int_\Omega (\operatorname{div}\mathbf{b})(u-v)^2\,\mathrm{d}\mathbf{x} + \int_{\Gamma_N} \kappa (u-v)^2\,\mathrm{d}s, \end{aligned}$$

the left-hand side of (2.44) transforms into the norm $|[u-v]|^2$. Now we use r_Ω and Green's first identity (A.5) to rewrite the equation

$$\begin{aligned} |[u-v]|^2 = \mathfrak{R}(v,\mathbf{y}) := &\int_\Omega r_\Omega(v,\mathbf{y})(u-v)\,\mathrm{d}\mathbf{x} + \int_\Omega \langle \mathbf{y} - \mathbf{A}\nabla v, \nabla(u-v)\rangle\,\mathrm{d}\mathbf{x} \\ &+ \int_{\Gamma_N} (g_N - \langle \mathbf{y},\mathbf{n}\rangle)(u-v)\,\mathrm{d}s, \end{aligned} \tag{2.45}$$

where $\mathbf{y}$ is an arbitrary function in

$$H_{\Gamma_N}(\Omega, \operatorname{div}) := \left\{\mathbf{y} \in H(\Omega, \operatorname{div}) \mid \langle \mathbf{y},\mathbf{n}\rangle \in L^2(\Gamma_N)\right\}.$$

The first term on the right-hand side of (2.45) can be estimated as

$$\begin{aligned} \left|\int_\Omega r_\Omega(v,\mathbf{y})(u-v)\,\mathrm{d}\mathbf{x}\right| &\le \|r_\Omega(v,\mathbf{y})\|_{L^2(\Omega)}\|u-v\|_{L^2(\Omega)} \\ &\le C_{F\Omega}\|r_\Omega(v,\mathbf{y})\|_{L^2(\Omega)}\|\nabla(u-v)\|_{L^2(\Omega)} \\ &\le \frac{C_{F\Omega}}{\sqrt{\alpha^{\mathrm{ell}}}}\|r_\Omega(v,\mathbf{y})\|_{L^2(\Omega)}\|\nabla(u-v)\|_{\mathbf{A}}. \end{aligned}$$

Similarly, we proceed for the third term,

$$\left|\int_{\Gamma_N} (g_N - \langle \mathbf{y},\mathbf{n}\rangle)(u-v)\,\mathrm{d}s\right| \le \frac{C_{T\Gamma_N}}{\sqrt{\alpha^{\mathrm{ell}}}}\|g_N - \langle \mathbf{y},\mathbf{n}\rangle\|_{L^2(\Gamma_N)}\|\nabla(u-v)\|_{\mathbf{A}}.$$

This leads to

$$\begin{aligned}\mathfrak{R}(v,y) &\le \|\mathbf{y}-\mathbf{A}\nabla v\|_{\mathbf{A}^{-1}} \|\nabla(u-v)\|_{\mathbf{A}} \\ &\quad + \left(\frac{C_{F\Omega}}{\sqrt{\alpha^{\mathrm{ell}}}} \|r_\Omega(v,\mathbf{y})\|_{L^2(\Omega)} + \frac{C_{T\Gamma_N}}{\sqrt{\alpha^{\mathrm{ell}}}} \|g_N - \langle \mathbf{y},\mathbf{n}\rangle\|_{L^2(\Gamma_N)}\right) \|\nabla(u-v)\|_{\mathbf{A}} \\ &\le \|\mathbf{y}-\mathbf{A}\nabla v\|_{\mathbf{A}^{-1}} |[u-v]| \\ &\quad + \frac{1}{\sqrt{\alpha^{\mathrm{ell}}}} \left(C_{F\Omega} \|r_\Omega(v,\mathbf{y})\|_{L^2(\Omega)} + C_{T\Gamma_N} \|g_N - \langle \mathbf{y},\mathbf{n}\rangle\|_{L^2(\Gamma_N)}\right) |[u-v]|. \end{aligned} \tag{2.46}$$

The estimate of the theorem follows from (2.45) and (2.46). □

For the practical computation, as explained in Chapter 5, it is more convenient to consider the majorant squared and to plug in constants $\beta,\gamma>0$ with Young's inequality (A.1), i.e.:

$$\begin{aligned}\mathcal{M}^2(v,\mathbf{y};\beta,\gamma) &:= (1+\beta)\|\mathbf{y}-\mathbf{A}\nabla v\|^2_{\mathbf{A}^{-1}} \\ &\quad + \frac{1+\beta}{\beta}\frac{1}{\alpha^{\mathrm{ell}}}\left((1+\gamma)C^2_{F\Omega}\|r_\Omega(v,\mathbf{y})\|^2_{L^2(\Omega)} + \frac{1+\gamma}{\gamma}C^2_{T\Gamma_N}\|g_N-\langle \mathbf{y},\mathbf{n}\rangle\|^2_{L^2(\Gamma_N)}\right).\end{aligned} \tag{2.47}$$

In the case of Dirichlet boundary condition on the whole boundary, we have

$$\mathcal{M}^2(v,\mathbf{y};\beta) := (1+\beta)\|\mathbf{y}-\mathbf{A}\nabla v\|^2_{\mathbf{A}^{-1}} + \frac{1+\beta}{\beta}\frac{C^2_{F\Omega}}{\alpha^{\mathrm{ell}}}\|r_\Omega(v,\mathbf{y})\|^2_{L^2(\Omega)}. \tag{2.48}$$

Remark 2.38. *The goal is to compute efficiently a sequence of approximate solutions converging to the exact solution and to verify the accuracy of those approximations reliably. An error majorant should give us a fully computable, guaranteed upper bound, i.e.,*

$$\|u-v\|_V \le \mathcal{M}(v,\mathbf{y},\mathcal{D}),$$

for the corresponding energy space V and the data $\mathcal{D}$ (coefficients, domain, initial conditions, etc.) of the boundary value problem. For example for the majorant (2.48), we can see that the closer $\mathbf{y}$ lies to the exact flux $\mathbf{A}\nabla u$, the sharper is the estimate. Furthermore, this majorant always yields a guaranteed upper bound for the error.
The flux ∇u or $\mathbf{A}\nabla u$ is often called the dual variable, therefore we sometimes denote

$$\mathcal{M}_{\mathrm{D}} := \|\mathbf{y}-\mathbf{A}\nabla v\|_{\mathbf{A}^{-1}}. \tag{2.49}$$

The equation $r_\Omega(v,\mathbf{y}) = f - \langle \mathbf{b},\nabla v\rangle - cv + \operatorname{div}\mathbf{y} = 0$ is a simple equilibration or balance equation, therefore we sometimes denote

$$\mathcal{M}_{\mathrm{Eq}} := \|r_\Omega(v,\mathbf{y})\|_{L^2(\Omega)}. \tag{2.50}$$

For a good choice of $\mathbf{y}$, which we explain in the next section, the value $\mathcal{M}_{\mathrm{D}}$ is an accurate error indicator and the term $\mathcal{M}_{\mathrm{Eq}}$ ensures reliability, where $\mathcal{M}_{\mathrm{Eq}} \approx 0$.

2.5 Gradient Recovery

Assume that we can compute the approximate solution $v_h \in V_h$, where $V_h \subset H^1(\Omega)$ is the finite element space, for example generated by continuous piecewise linear elements on a given triangulation. Note that v_h may differ from the Galerkin approximation u_h due to roundoff or integration errors. With v_h we can construct an approximation of the flux

$$p_h := \mathbf{A}\nabla v_h \in L^2\left(\Omega,\mathbb{R}^2\right).$$

In the a posteriori error estimates, e.g. (2.47), where we will have to plug in $\mathbf{y}=p_h$, we need some regularity of the flux, which is given by a priori properties. More precisely, p_h has to be in $H(\Omega,\mathrm{div})$. This can be achieved by a post-processing operator

$$G_h : L^2\left(\Omega,\mathbb{R}^2\right) \to H^1\left(\Omega,\mathbb{R}^2\right) \subset H(\Omega,\mathrm{div}).$$

A suitable recovery procedure should be rather inexpensive and should give an approximation of the flux which is closer to the original flux. Below we will discuss a global gradient recovery procedure, for other methods we refer to e.g. [26] and [24].
In the following we will use the gradient recovery strategy from [6]. This means, that we first use the L^2-projection operator Q_h, which already projects on the space of continuous piecewise linear polynomials, and then we use additionally a smoothing operator S, i.e., $G_h(\nabla v_h) := SQ_h\nabla v_h$.
Consider the discrete L^2-projection operator:

$$Q_h : L^2\left(\Omega, \mathbb{R}^2\right) \to V_h^2 \subset H^1\left(\Omega, \mathbb{R}^2\right),$$

defined by

$$\left(Q_h\nabla v_h, \mathbf{w}_h\right)_{L^2(\Omega)} = \left(\nabla v_h, \mathbf{w}_h\right)_{L^2(\Omega)}, \quad \forall \mathbf{w}_h \in L^2\left(\Omega, \mathbb{R}^2\right). \tag{2.51}$$

The smoothing operator S is constructed as follows. Consider the bilinear form $b : H^1(\Omega)\times H^1(\Omega) \to \mathbb{R}$ defined by

$$b(u, v) := \left(\nabla u, \nabla v\right)_{L^2(\Omega)} + \left(u, v\right)_{L^2(\Omega)}.$$

By the definition

$$\left(A_h u_h, v_h\right)_{L^2(\Omega)} := b(u_h, v_h) \quad \forall u_h, v_h \in V_h,$$

we have the discrete operator $A_h : v_h \to V_h$. The operator A_h is symmetric and positive definite on V_h and we set $\lambda \equiv \rho(A_h)$, where $\rho(A_h)$ denotes the spectral radius of A_h. With the considerations from Section A.3 and with the inverse estimate from Remark 2.30, we know

$$\rho(A_h) \le \sup_{v_h\in H_0^1(\Omega)\setminus\{0\}} \frac{\|\nabla v_h\|_{L^2(\Omega)}^2}{\|v_h\|_{L^2(\Omega)}^2} \le \sup_{v_h\in H_0^1(\Omega)\setminus\{0\}} \frac{c_{\text{inv}}^2 h^{-2}\|v_h\|_{L^2(\Omega)}^2}{\|v_h\|_{L^2(\Omega)}^2}.$$

Therefore, we have

$$\lambda = \rho(A_h) \le c_{\text{inv}}^2 h^{-2}.$$

Then, the smoothing operator S is defined by

$$S := I - \lambda^{-1} A_h,$$

where I is the identity operator. More generally, we can apply several smoothing steps by applying the operator S m-times for a positive integer m, i.e.,

$$G_h(\nabla v_h) := S^m Q_h \nabla v_h. \tag{2.52}$$

The implementation of the L^2-projection and the operator S will be explained in Chapter 5.
In [5], a gradient recovery scheme without the smoothing operator S is considered, hence $G_h\nabla u_h = Q_h\nabla u_h$, and it is shown that this gives a superconvergent approximation to ∇u, i.e., it converges with a rate higher than the expected a priori rate; for more theory about superconvergence see, e.g., [34]. The arguments base mainly on the geometry of the underlying triangular mesh. In particular, it is shown that if the triangulation $\mathcal{T}_h$ is $O\left(h^{2\sigma}\right)$ irregular and $u \in W^{3,\infty}(\Omega)$, it holds:

$$\|\nabla u - Q_h\nabla u_h\|_{L^2(\Omega)} \lesssim h^{1+\min(1,\sigma)} |\log(h)|^{1/2} \|u\|_{W^{3,\infty}(\Omega)}. \tag{2.53}$$

The $O\left(h^{2\sigma}\right)$ irregular property means, that the estimate holds for quasi-uniform[2] meshes, where an $O\left(h^2\right)$ approximate parallelogram property is satisfied for pairs of adjacent triangles in most parts of Ω, except for a region of size $O\left(h^{2\sigma}\right)$, see [5] for more details. If $\sigma > 0$ becomes very close to zero, estimate (2.53) no longer gives superconvergence. The idea is, that this is due to high frequency errors which is well studied in multilevel methods, thus they propose to use an appropriate multigrid-like smoothing operator S.
In [6], superconvergence is developed for general unstructured, but shape regular meshes, where the idea of smoothing iteration of the multigrid method is used. Hence, $G_h\nabla u_h = S^m Q_h\nabla u_h$, with S an

[2] Quasi-uniformity is a weaker requirement than uniformity, but a stronger requirement than shape regularity.

appropriate smoothing operator and m a positive integer, preferably small. The number σ measures in some sense the extent to which the $O\left(h^2\right)$ approximate parallelogram property is violated. If σ is sufficiently large, $m = 0$ can be chosen, since the L^2-projection is sufficient, $m > 0$ is needed for $\sigma \approx 0$. The following main result is presented: If the triangulation $\mathcal{T}_h$ is $O\left(h^{2\sigma}\right)$ irregular and $u \in W^{3,\infty}(\Omega)$, it holds:

$$\|\nabla u - S^m Q_h \nabla u_h\|_{L^2(\Omega)} \lesssim h\left\{\min\left(h^{\min(1,\sigma)}|\log h|, \varepsilon_m\right) + mh^{1/2}\right\}\|u\|_{W^{3,\infty}(\Omega)}, \tag{2.54}$$

where

$$\varepsilon_m = \begin{cases} \kappa^{\alpha/2} f\left(m, \alpha/2\right) \lesssim m^{-\alpha/2} & \text{for } m > (\kappa-1)\alpha/2, \\ \left[(\kappa-1)/\kappa\right]^m & \text{for } m \leq (\kappa-1)\alpha/2, \end{cases}$$

$1/2 < \alpha < 1$, f is the usual multigrid convergence function $f(\alpha,\beta) = \frac{\alpha^\alpha \beta^\beta}{(\alpha+\beta)^{\alpha+\beta}}$ for $\alpha, \beta > 0$, and $\kappa = O(1)$. The term $\left(1-\kappa^{-1}\right)^m$ illustrates the well-known effectiveness of a few smoothing steps and is reminiscent of terms arising in connection with multigrid convergence analysis.

3 Homogenization

In this chapter we explain the homogenization theory as it is studied, e.g., in [13], [22] and [8]. One application is the study of composite materials, which have certain better properties for applications such as lightweight constructions. They are characterised by a main homogeneous material containing small heterogeneities, which can be modelled by a periodic structure. A partial differential equation describing this behaviour would have rapidly oscillating coefficients. To solve this problem accurately, one would have to resolve the oscillations, which is numerically not feasible. Asymptotically, we can think of a macroscopic scale, describing the global behaviour, and of a microscopic scale, describing the small heterogeneities. This approach leads to problems without oscillations that can be solved numerically and will be described in detail in the following.

3.1 Introduction to Homogenization

We consider the homogenization of an elliptic boundary value problem within a periodic structure with the following setting (see [28]). Let $\Omega \subset \mathbb{R}^d$ be an open, bounded, convex and polygonal domain with Lipschitz boundary $\Gamma := \partial\Omega$. The periodic structure is defined with repeating elements $\Omega = \bigcup_{\mathbf{i}} \Pi_{\mathbf{i}}^{\varepsilon}$, where the general **cell** is defined as

$$\Pi_{\mathbf{i}}^{\varepsilon} = \mathbf{x_i} + \varepsilon\hat{\Pi} = \left\{ \mathbf{x} \in \mathbb{R}^d \,\middle|\, \frac{\mathbf{x} - \mathbf{x_i}}{\varepsilon} \in \widehat{\Pi} \right\},$$

with the reference point $\mathbf{x_i}$ of $\Pi_{\mathbf{i}}^{\varepsilon}$, see Figure 3.1. The global coordinates are $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{i} = (i_1, \ldots, i_d)$ are multi-indices and $\bigcup_{\mathbf{i}}$ means the union over all cells. The amount of cells is bounded from above by $c\varepsilon^{-2}$, where $c = O(1)$. The local coordinates $\mathbf{y} \in \mathbb{R}^d$ are used in the basic cell, see Figure 3.1. The global and local coordinates are connected by the following relation

$$\mathbf{y} = \frac{\mathbf{x} - \mathbf{x_i}}{\varepsilon} \in \widehat{\Pi} \qquad \forall \mathbf{x} \in \Pi_{\mathbf{i}}^{\varepsilon}, \forall \mathbf{i}.$$

Figure 3.1: Periodic structure of Ω and the basic cell $\widehat{\Pi}$ in two dimensions.

On the cell $\widehat{\Pi}$ we consider the coefficients $\widehat{\mathbf{A}} \in L^{\infty}\left(\widehat{\Pi}, \mathbb{R}^{d\times d}_{\mathrm{sym}}\right)$, $\widehat{\mathbf{b}} \in L^{\infty}\left(\widehat{\Pi}, \mathbb{R}^d\right)$, with $\operatorname{div}\widehat{\mathbf{b}} \in L^{\infty}\left(\widehat{\Pi}, \mathbb{R}\right)$, and $\hat{c} \in L^{\infty}\left(\widehat{\Pi}, \mathbb{R}\right)$. We assume that for the symmetric coefficient matrices it holds

$$\widehat{\alpha}^{\mathrm{ell}} \|\xi\|_2^2 \le \left\langle \widehat{\mathbf{A}}(\mathbf{y})\xi, \xi \right\rangle \le \widehat{\alpha}^{\mathrm{cont}} \|\xi\|_2^2 \qquad \forall \xi \in \mathbb{R}^d, \forall \mathbf{y} \in \widehat{\Pi}, \tag{3.1}$$

where $0 < \widehat{\alpha}^{\mathrm{ell}} \le \widehat{\alpha}^{\mathrm{cont}} < \infty$. Further, we assume

$$\hat{c}(\mathbf{y}) \ge \widehat{\alpha} > 0, \quad \forall \mathbf{y} \in \widehat{\Pi} \quad \text{and} \quad -\frac{1}{2}\operatorname{div}\widehat{\mathbf{b}} + \hat{c} \ge 0. \tag{3.2}$$

Then, the global coefficients are defined as

$$\mathbf{A}_\varepsilon(\mathbf{x}) := \widehat{\mathbf{A}}\left(\frac{\mathbf{x}-\mathbf{x_i}}{\varepsilon}\right), \quad \mathbf{b}_\varepsilon(\mathbf{x}) := \widehat{\mathbf{b}}\left(\frac{\mathbf{x}-\mathbf{x_i}}{\varepsilon}\right), \quad c_\varepsilon(\mathbf{x}) := \hat{c}\left(\frac{\mathbf{x}-\mathbf{x_i}}{\varepsilon}\right), \quad \forall \mathbf{x} \in \Pi_{\mathbf{i}}^\varepsilon, \forall \mathbf{i} \tag{3.3}$$

and $\widehat{\mathbf{A}}_{ij}$, $\hat{b}_i$ and $\hat{c}$ are all $\widehat{\Pi}$-periodic. The matrix $\mathbf{A}_\varepsilon$ is also symmetric and satisfies similar inequalities like (3.1) with constants $\alpha_\varepsilon^{\text{ell}}$ and $\alpha_\varepsilon^{\text{cont}}$.
The norm $\|\mathbf{A}\|_{p,\Omega}$ and the spectral radius $\rho_\Omega(\mathbf{A})$ for a general matrix function $\mathbf{A} \in L^\infty\left(\Omega, \mathbb{R}^{d\times d}\right)$ are defined in Section A.1. Note that it holds $\left(\widehat{\alpha}^{\text{ell}}\right)^{-1} = \rho_{\widehat{\Pi}}\left(\widehat{\mathbf{A}}^{-1}\right)$ and $\widehat{\alpha}^{\text{cont}} = \rho_{\widehat{\Pi}}\left(\widehat{\mathbf{A}}\right)$.
For $f \in L^2(\Omega)$ we consider the elliptic partial differential equation of second order

$$-\operatorname{div}\left(\mathbf{A}_\varepsilon \nabla u_\varepsilon\right) + \langle \mathbf{b}_\varepsilon, \nabla u_\varepsilon \rangle + c_\varepsilon u_\varepsilon = f \quad \text{in } \Pi_{\mathbf{i}}^\varepsilon, \quad \forall \mathbf{i}, \tag{3.4}$$

with Dirichlet boundary condition $u_\varepsilon = g$ on Γ, $g \in L^2(\Gamma)$. Hence, we are looking for the solution $u_\varepsilon \in H^1(\Omega)$ such that

$$a_\varepsilon(u_\varepsilon, v) := \int_\Omega \left\{\langle \mathbf{A}_\varepsilon \nabla u_\varepsilon, \nabla v\rangle + \langle \mathbf{b}_\varepsilon, \nabla u_\varepsilon\rangle v + c_\varepsilon u_\varepsilon v\right\} \mathrm{d}\mathbf{x} = \int_\Omega f v \,\mathrm{d}\mathbf{x} =: l(v) \quad \forall v \in H_0^1(\Omega),$$

for any $\varepsilon > 0$. Proposition 2.3 gives the existence and uniqueness of $u_\varepsilon \in H^1(\Omega)$.
The theory of homogenization problems is well studied. In particular, it is possible to find a two scale approximation

$$u_\varepsilon^1(\mathbf{x}) = u_0(\mathbf{x}) + \varepsilon u_1(\mathbf{x}, \mathbf{y}), \quad \forall \mathbf{x} \in \Omega, \forall \mathbf{y} \in \widehat{\Pi},$$

of u_ε, that fulfils the a priori error estimate

$$\left\|u_\varepsilon - u_\varepsilon^1\right\|_{H^1(\Omega)} \le c\sqrt{\varepsilon}$$

under certain assumptions on f, u_0, u_1, Ω, $\mathbf{A}_\varepsilon$, $\mathbf{b}_\varepsilon$ and c_ε. In Section 3.2 we will investigate the two scale approximation and derive u_0 and u_1. In Section 3.3 we will prove the a priori error estimate. In Section 3.4 we will verify several properties of the homogenized coefficients.

3.2 Two Scale Approximation

In this context we think of $\mathbf{x}$ and $\mathbf{y} = \varepsilon^{-1}\mathbf{x}$ as two independent variables, where $\mathbf{x} \in \Omega$ is the macroscopic and $\mathbf{y} \in \widehat{\Pi}$ is the microscopic scale. Therefore we use an asymptotic expansion (see e.g. [8, Chapter 13], [13, Chapter 7] or [22, Section 1.4]) for u_ε:

$$u_\varepsilon^1(\mathbf{x}) = u_0(\mathbf{x}, \mathbf{y}) + \varepsilon u_1(\mathbf{x}, \mathbf{y}), \quad \forall \mathbf{x} \in \Omega, \forall \mathbf{y} \in \widehat{\Pi}, \tag{3.5}$$

where $u_0(\mathbf{x}, \mathbf{y}), u_1(\mathbf{x}, \mathbf{y})$ are $\widehat{\Pi}$-periodic functions in $\mathbf{y}$. Consider a function φ which depends on both variables $\varphi = \varphi(\mathbf{x}, \mathbf{y})$ and the corresponding function φ_ε which depends only on one variable denoted by

$$\varphi_\varepsilon(\mathbf{x}) = \varphi\left(\mathbf{x}, \frac{\mathbf{x}}{\varepsilon}\right).$$

We can now apply the derivative with respect to x_i and get the following:

$$\partial_{x_i}\varphi_\varepsilon(\mathbf{x}) = \partial_{x_i}\varphi\left(\mathbf{x}, \frac{\mathbf{x}}{\varepsilon}\right) + \frac{1}{\varepsilon}\partial_{y_i}\varphi\left(\mathbf{x}, \frac{\mathbf{x}}{\varepsilon}\right).$$

With these considerations we can rewrite the left-hand side of the equation (3.4):

$$\begin{aligned} -\operatorname{div}\left(\mathbf{A}_\varepsilon \nabla u_\varepsilon^1\right) + \left\langle \mathbf{b}_\varepsilon, \nabla u_\varepsilon^1\right\rangle + c_\varepsilon u_\varepsilon^1 &= \left(-\sum_{i,j=1}^d \partial_{x_i}\left(\hat{a}_{i,j}(\mathbf{y})\partial_{x_i}\right) + \hat{c}(\mathbf{y})\right) u_\varepsilon^1(\mathbf{x}) + \sum_{i=1}^d \hat{b}_i(\mathbf{y})\partial_{x_i} u_\varepsilon^1(\mathbf{x}) \\ &=: \mathcal{A}_\varepsilon u_\varepsilon^1 + \mathcal{B}_\varepsilon u_\varepsilon^1. \end{aligned} \tag{3.6}$$

The operators $\mathcal{A}_\varepsilon$ and $\mathcal{B}_\varepsilon$ have the form

$$\begin{aligned}\mathcal{A}_\varepsilon &= -\sum_{i,j=1}^{d}\left(\partial_{x_i}+\frac{1}{\varepsilon}\partial_{y_i}\right)\hat{a}_{i,j}\left(\partial_{x_j}+\frac{1}{\varepsilon}\partial_{y_j}\right)+\hat{c}\\ &=\sum_{i,j=1}^{d}\left(-\partial_{x_i}\left(\hat{a}_{i,j}\partial_{x_j}\right)+\varepsilon^{-1}\left(-\partial_{y_i}\left(\hat{a}_{i,j}\partial_{x_j}\right)-\partial_{x_i}\left(\hat{a}_{i,j}\partial_{y_j}\right)\right)+\varepsilon^{-2}\left(-\partial_{y_i}\left(\hat{a}_{i,j}\partial_{y_j}\right)\right)\right)+\hat{c}\\ &=:\mathcal{A}_2+\varepsilon^{-1}\mathcal{A}_1+\varepsilon^{-2}\mathcal{A}_0 \end{aligned} \tag{3.7}$$

and

$$\mathcal{B}_\varepsilon=\sum_{i=1}^{d}\hat{b}_i\left(\partial_{x_i}+\frac{1}{\varepsilon}\partial_{y_i}\right)=\sum_{i=1}^{d}\left(\hat{b}_i\partial_{x_i}+\varepsilon^{-1}\hat{b}_i\partial_{y_i}\right)=:\mathcal{B}_1+\varepsilon^{-1}\mathcal{B}_0, \tag{3.8}$$

with

$$\begin{aligned}\mathcal{A}_0&=-\sum_{i,j=1}^{d}\partial_{y_i}\left(\hat{a}_{i,j}\partial_{y_j}\right),\\ \mathcal{A}_1&=-\sum_{i,j=1}^{d}\partial_{y_i}\left(\hat{a}_{i,j}\partial_{x_j}\right)-\sum_{i,j=1}^{d}\partial_{x_i}\left(\hat{a}_{i,j}\partial_{y_j}\right), \qquad (3.9)\\ \mathcal{A}_2&=-\sum_{i,j=1}^{d}\partial_{x_i}\left(\hat{a}_{i,j}\partial_{x_j}\right)+\hat{c},\\ \mathcal{B}_0&=\sum_{i=1}^{d}\hat{b}_i\partial_{y_i},\quad \mathcal{B}_1=\sum_{i=1}^{d}\hat{b}_i\partial_{x_i}. \qquad (3.10)\end{aligned}$$

Inserting equation (3.5) into equation $\mathcal{A}_\varepsilon u_\varepsilon^1+\mathcal{B}_\varepsilon u_\varepsilon^1=f$ and using (3.7) and (3.8) gives:

$$\begin{aligned}f&=\left(\mathcal{A}_2+\varepsilon^{-1}\mathcal{A}_1+\varepsilon^{-2}\mathcal{A}_0\right)(u_0+\varepsilon u_1)+\left(\mathcal{B}_1+\varepsilon^{-1}\mathcal{B}_0\right)(u_0+\varepsilon u_1)\\ &=\varepsilon^{-2}\mathcal{A}_0u_0+\varepsilon^{-1}\left(\mathcal{A}_0u_1+\left(\mathcal{A}_1+\mathcal{B}_0\right)u_0\right)+\left(\left(\mathcal{A}_1+\mathcal{B}_0\right)u_1+\left(\mathcal{A}_2+\mathcal{B}_1\right)u_0\right)+\varepsilon\left(\mathcal{A}_2+\mathcal{B}_1\right)u_1,\end{aligned}$$

which results in a system of equations:

$$\mathcal{A}_0u_0=0, \tag{3.11}$$

$$\mathcal{A}_0u_1=-\left(\mathcal{A}_1+\mathcal{B}_0\right)u_0, \tag{3.12}$$

$$\left(\mathcal{A}_1+\mathcal{B}_0\right)u_1+\left(\mathcal{A}_2+\mathcal{B}_1\right)u_0=f, \tag{3.13}$$

$$\left(\mathcal{A}_2+\mathcal{B}_1\right)u_1=0. \tag{3.14}$$

Notice that the equations (3.11), (3.12) and (3.14) are boundary value problems with periodic boundary conditions and (3.13) inherits the boundary condition from the original problem. In Subsection 2.2.3 we have seen, that a solution of a periodic boundary value problem can either be stated in the sense of a class of equivalence (2.22) or as a function with zero mean value (2.24). For each problem we will choose the formulation which is more suitable.
Observe that $\mathcal{A}_0u=-\operatorname{div}_{\mathbf{y}}\left(\widehat{\mathbf{A}}\nabla_{\mathbf{y}}u\right)$ and that $\mathbf{x}$ is a parameter. Let us consider the equations separately. Since the first equation (3.11) contains only one unknown u_0 and if we know u_0, we can derive u_1 with equation (3.12).
The variational formulation of equation (3.11) is:

$$\text{Find } \dot{u}_0\in W_{\text{per}}(\widehat{\Pi}) \text{ such that } \int_{\widehat{\Pi}}\left\langle\widehat{\mathbf{A}}\nabla_{\mathbf{y}}u_0,\nabla_{\mathbf{y}}v\right\rangle\mathrm{d}\mathbf{y}=0\quad\forall v\in W_{\text{per}}\left(\widehat{\Pi}\right),\forall u_0\in\dot{u}_0,\forall v\in\dot{v}.$$

We have seen that functions of the quotient space $W_{\text{per}}\left(\widehat{\Pi}\right)$ are defined up to an additive constant. Hence, we obtain $\dot{u}_0=\dot{0}$ in $W_{\text{per}}\left(\widehat{\Pi}\right)$ as the unique solution and since $\mathbf{x}$ is a parameter, we get

$$u_0(\mathbf{x},\mathbf{y})=u_0(\mathbf{x}),\quad\forall u_0\in\dot{u}_0. \tag{3.15}$$

According to [13], we expect u_0 to be the **homogenized solution**, since u_0 does not depend on the rapidly oscillating scale $\mathbf{x}/\varepsilon$.
Using (3.15), and therefore $\partial_{y_i} u_0 = 0$, equation (3.12) becomes

$$\mathcal{A}_0 u_1 = \sum_{i,j=1}^{d} \left(\partial_{y_i} \hat{a}_{i,j}\right) \left(\partial_{x_j} u_0\right).$$

Hence, the variational formulation for this equation is:

$$\text{Find } \dot{u}_1 \in W_{\text{per}}\left(\widehat{\Pi}\right) \text{ such that } \int_{\hat{\Pi}} \left\langle \widehat{\mathbf{A}} \nabla_{\mathbf{y}} u_1, \nabla_{\mathbf{y}} v \right\rangle \mathrm{d}\mathbf{y} = l(v) \quad \forall v \in W_{\text{per}}\left(\widehat{\Pi}\right), \forall u_1 \in \dot{u}_1, \forall v \in \dot{v}, \tag{3.16}$$

where the functional l is given by

$$\begin{aligned} l(v) &= \sum_{i,j=1}^{d} \left(\partial_{x_j} u_0\right) \int_{\hat{\Pi}} \left(\partial_{y_i} \hat{a}_{i,j}\right) v \, \mathrm{d}\mathbf{y} \\ &= -\sum_{i,j=1}^{d} \left(\partial_{x_j} u_0\right) \int_{\hat{\Pi}} \hat{a}_{i,j} \partial_{y_i} v \, \mathrm{d}\mathbf{y}, \quad \forall v \in W_{\text{per}}\left(\widehat{\Pi}\right), \forall v \in \dot{v}. \end{aligned}$$

For two elements in the equivalence class, $v_1, v_2 \in \dot{v}$, it holds

$$\partial_{y_i} v_1 = \partial_{y_i} v_2$$

and therefore

$$l(v_1) = l(v_2).$$

With Proposition 2.12, this defines l as an element of $\left(W_{\text{per}}\left(\widehat{\Pi}\right)\right)'$ and hence the definition of l makes sense. The variational formulation (3.16) can be stated in the following equivalent form, see Subsection 2.2.3:

$$\text{Find } u_1 \in W_{\text{per}}\left(\widehat{\Pi}\right) \text{ such that } \int_{\hat{\Pi}} \left\langle \widehat{\mathbf{A}} \nabla_{\mathbf{y}} u_1, \nabla_{\mathbf{y}} v \right\rangle \mathrm{d}\mathbf{y} = l(v) \quad \forall v \in W_{\text{per}}\left(\widehat{\Pi}\right), \tag{3.17}$$

where $W_{\text{per}}\left(\widehat{\Pi}\right)$ is understood in the sense of zero mean value as in (2.25). Proposition 2.13 gives the existence and uniqueness of a solution $u_1 \in W_{\text{per}}\left(\widehat{\Pi}\right)$ of equation (3.17). Notice the linearity of $\mathcal{A}_0$ and that $\mathcal{A}_0$ is independent of $\mathbf{x}$ and $\partial_{x_j} u_0$ is independent of $\mathbf{y}$. Therefore, for more details see [13], any solution u_1 has the form

$$u_1(\mathbf{x}, \mathbf{y}) = -\sum_{j=1}^{d} \widehat{N}_j(\mathbf{y}) \, \partial_{x_j} u_0, \quad \text{in } W_{\text{per}}\left(\widehat{\Pi}\right). \tag{3.18}$$

For all $i = 1, \ldots, d$ one can easily check that $\widehat{N}_j \in W_{\text{per}}\left(\widehat{\Pi}\right)$ is the solution of the partial differential equation

$$\begin{aligned} \mathcal{A}_0 \widehat{N}_j &= -\sum_{i=1}^{d} \partial_{y_i} \hat{a}_{i,j} \quad \text{in } \widehat{\Pi}, \\ \widehat{N}_j &\quad \widehat{\Pi}\text{ – periodic }, \\ \left\langle \widehat{N}_j \right\rangle_{\widehat{\Pi}} &= 0, \end{aligned} \tag{3.19}$$

or in weak formulation:

$$\text{Find } \widehat{N}_j \in W_{\text{per}}\left(\widehat{\Pi}\right) \text{ such that } \int_{\hat{\Pi}} \langle \widehat{\mathbf{A}} \nabla_{\mathbf{y}} \widehat{N}_j, \nabla_{\mathbf{y}} v \rangle \, \mathrm{d}\mathbf{y} = \sum_{i=1}^{d} \int_{\hat{\Pi}} \hat{a}_{i,j} \partial_{y_i} v \, \mathrm{d}\mathbf{y} \quad \forall v \in W_{\text{per}}\left(\widehat{\Pi}\right).$$

Proposition 2.13 gives also the existence and uniqueness of $\widehat{N}_j \in W_{\text{per}}\left(\widehat{\Pi}\right)$, for all $i = 1, \ldots, d$. Equation (3.19) is equivalent to the following equation, which we will call the **auxiliary cell problem**:

$$\begin{aligned} \operatorname{div}\left(\widehat{\mathbf{A}}(\mathbf{y}) \nabla \widehat{N}_j(\mathbf{y})\right) &= \operatorname{div}\left(\widehat{\mathbf{A}}_j(\mathbf{y})\right) \quad \text{in } \widehat{\Pi}, \\ \widehat{N}_j &\quad \widehat{\Pi}\text{ – periodic }, \\ \left\langle \widehat{N}_j \right\rangle_{\widehat{\Pi}} &= 0, \end{aligned} \tag{3.20}$$

where $\widehat{\mathbf{A}}_j$ is the j-th column of matrix $\widehat{\mathbf{A}}$.
Now we turn our attention to equation (3.13), the left-hand side states:

$$\begin{aligned}(\mathcal{A}_1+\mathcal{B}_0)\,u_1+(\mathcal{A}_2+\mathcal{B}_1)\,u_0 &= \left(-\sum_{i,j=1}^{d}\partial_{y_i}\left(\hat{a}_{i,j}\partial_{x_j}\right)-\sum_{i,j=1}^{d}\partial_{x_i}\left(\hat{a}_{i,j}\partial_{y_j}\right)+\sum_{i=1}^{d}\hat{b}_i\partial_{y_i}\right)u_1\\ &\quad+\left(-\sum_{i,j=1}^{d}\partial_{x_i}\left(\hat{a}_{i,j}\partial_{x_j}\right)+\hat{c}+\sum_{i=1}^{d}\hat{b}_i\partial_{x_i}\right)u_0.\end{aligned}$$

Using equations (3.15) and (3.18), and that $\hat{a}_{i,j}(\mathbf{y})$ depends only on $\mathbf{y}$, this is equivalent to:

$$\begin{aligned}(\mathcal{A}_1+\mathcal{B}_0)\,u_1+(\mathcal{A}_2+\mathcal{B}_1)\,u_0 &= \sum_{k=1}^{d}\left(\sum_{i,j=1}^{d}\partial_{y_i}\left(\hat{a}_{i,j}\widehat{N}_k\right)\partial_{x_j}\partial_{x_k}u_0+\sum_{i,j=1}^{d}\hat{a}_{i,j}\partial_{y_j}\widehat{N}_k\partial_{x_i}\partial_{x_k}u_0\right.\\ &\quad\left.-\sum_{i=1}^{d}\hat{b}_i\partial_{y_i}\widehat{N}_k\partial_{x_k}u_0\right)-\sum_{i,j=1}^{d}\hat{a}_{i,j}\partial_{x_i}\partial_{x_j}u_0+\hat{c}u_0+\sum_{i=1}^{d}\hat{b}_i\partial_{x_i}u_0\\ &= \sum_{k,i,j=1}^{d}\partial_{y_i}\left(\hat{a}_{i,j}\widehat{N}_k\right)\partial_{x_j}\partial_{x_k}u_0+\sum_{i,k=1}^{d}\left(\sum_{j=1}^{d}\hat{a}_{i,j}\partial_{y_j}\widehat{N}_k-\hat{a}_{i,k}\right)\partial_{x_i}\partial_{x_k}u_0\\ &\quad+\sum_{k=1}^{d}\left(\hat{b}_k-\sum_{i=1}^{d}\hat{b}_i\partial_{y_i}\widehat{N}_k\right)\partial_{x_k}u_0+\hat{c}u_0.\end{aligned}$$

Taking the mean value of this, we get:

$$\begin{aligned}\left\langle(\mathcal{A}_1+\mathcal{B}_0)\,u_1+(\mathcal{A}_2+\mathcal{B}_1)\,u_0\right\rangle_{\widehat{\Pi}} &= -\frac{1}{|\widehat{\Pi}|}\sum_{i,k=1}^{d}\int_{\hat{\Pi}}\left(\hat{a}_{i,k}(\mathbf{y})-\sum_{j=1}^{d}\hat{a}_{i,j}(\mathbf{y})\partial_{y_j}\widehat{N}_k(\mathbf{y})\right)\mathrm{d}\mathbf{y}\,\partial_{x_i}\partial_{x_k}u_0(\mathbf{x})\\ &\quad+\frac{1}{|\widehat{\Pi}|}\sum_{k=1}^{d}\int_{\hat{\Pi}}\left(\hat{b}_k(\mathbf{y})-\sum_{i=1}^{d}\hat{b}_i(\mathbf{y})\partial_{y_i}\widehat{N}_k(\mathbf{y})\right)\mathrm{d}\mathbf{y}\,\partial_{x_k}u_0(\mathbf{x})\\ &\quad+\frac{1}{|\widehat{\Pi}|}\int_{\hat{\Pi}}\hat{c}(\mathbf{y})\,\mathrm{d}\mathbf{y}\,u_0(\mathbf{x})\\ &=: -\sum_{i,k=1}^{d}\left(\mathbf{A}_0\right)_{i,k}\partial_{x_i}\partial_{x_k}u_0(\mathbf{x})+\sum_{k=1}^{d}\left(B_0\right)_k\partial_{x_k}u_0(\mathbf{x})+c_0u_0(\mathbf{x}).\end{aligned}$$

Remembering the right-hand side of (3.13), that it depends only on $\mathbf{x}$ and taking its mean value, we get the following **homogenized boundary value problem**:

$$-\sum_{i,k=1}^{d}\left(\mathbf{A}_0\right)_{i,k}\partial_{x_i}\partial_{x_k}u_0(\mathbf{x})+\sum_{k=1}^{d}\left(B_0\right)_k\partial_{x_k}u_0(\mathbf{x})+c_0u_0(\mathbf{x})=f\quad\text{in }\Omega,\tag{3.21}$$

with $u_0 = g$ on Γ. Proposition 2.3 gives, under conditions on $\mathbf{A}_0$, B_0 and c_0, which we will show in Section 3.4, the existence and uniqueness of $u_0 \in H^1(\Omega)$. The coefficients can be rewritten in a more compact way:

$$\begin{aligned}\left(\mathbf{A}_0\right)_{i,k} &= \frac{1}{|\widehat{\Pi}|}\int_{\hat{\Pi}}\left(\hat{a}_{i,k}(\mathbf{y})-\sum_{j=1}^{d}\hat{a}_{i,j}(\mathbf{y})\partial_{y_j}\widehat{N}_k(\mathbf{y})\right)\mathrm{d}\mathbf{y}\\ &= \left\langle\hat{a}_{i,k}-\widehat{\mathbf{A}}_i\nabla\widehat{N}_k\right\rangle_{\widehat{\Pi}},\end{aligned}$$

where $\widehat{\mathbf{A}}_i$ is the i-th row of the matrix $\widehat{\mathbf{A}}$. Further, we get

$$\mathbf{A}_0=\left\langle\widehat{\mathbf{A}}\left(I-\left(\nabla\widehat{\mathbf{N}}\right)^{\mathsf{T}}\right)\right\rangle_{\widehat{\Pi}},\tag{3.22}$$

with $\widehat{\mathbf{N}} = \left(\widehat{N}_k\right)_{k=1}^{d}$ being a row vector. For example in two dimensions the gradient (Jacobian matrix) of $\widehat{\mathbf{N}}$ denotes

$$\nabla\widehat{\mathbf{N}}=\begin{pmatrix}\partial_{y_1}\widehat{N}_1 & \partial_{y_2}\widehat{N}_1\\ \partial_{y_1}\widehat{N}_2 & \partial_{y_2}\widehat{N}_2\end{pmatrix}.$$

Similarly, we get

$$\begin{aligned}(B_0)_k &= \frac{1}{|\widehat{\Pi}|}\int_{\hat{\Pi}}\left(\hat{b}_k(\mathbf{y}) - \sum_{i=1}^{d}\hat{b}_i(\mathbf{y})\partial_{y_i}\widehat{N}_k(\mathbf{y})\right)\mathrm{d}\mathbf{y}\\ &= \left\langle \hat{b}_k - \left(\nabla\widehat{N}_k\right)^{\mathsf{T}}\widehat{\mathbf{b}}\right\rangle_{\widehat{\Pi}}\end{aligned}$$

and by considering the definition of $\nabla\widehat{\mathbf{N}}$, we have for B_0:

$$B_0 = \left\langle \left(I - \nabla\widehat{\mathbf{N}}\right)\widehat{\mathbf{b}}\right\rangle_{\widehat{\Pi}}. \tag{3.23}$$

Note that we defined

$$c_0 = \langle \hat{c}\rangle_{\widehat{\Pi}}. \tag{3.24}$$

Observe that both $\mathbf{A}_0$ and B_0 are constant matrices of size $d\times d$ and $d\times 1$, respectively, therefore equation (3.21) is equivalent to

$$-\operatorname{div}\left(\mathbf{A}_0\nabla u_0\right) + \langle B_0, \nabla u_0\rangle + c_0 u_0 = f \quad \text{in } \Omega.$$

In conclusion, we get four computational steps to get the approximation u_ε^1 of u_ε:

1) Compute the solutions $\widehat{N}_j \in W_{\mathrm{per}}\left(\widehat{\Pi}\right)$, $j = 1,\dots,d$, of the cell problems

$$\begin{aligned}\operatorname{div}\left(\widehat{\mathbf{A}}\nabla\widehat{N}_j\right) &= \operatorname{div}\left(\widehat{\mathbf{A}}_j\right) \quad \text{in } \widehat{\Pi},\\ \widehat{N}_j &\quad \widehat{\Pi}\text{ – periodic },\\ \left\langle\widehat{N}_j\right\rangle_{\widehat{\Pi}} &= 0.\end{aligned} \tag{3.25}$$

2) Compute the homogenized coefficients:

$$\begin{aligned}\mathbf{A}_0 &= \left\langle\widehat{\mathbf{A}}\left(I - \left(\nabla\widehat{\mathbf{N}}\right)^{\mathsf{T}}\right)\right\rangle_{\widehat{\Pi}},\\ B_0 &= \left\langle\left(I - \nabla\widehat{\mathbf{N}}\right)\widehat{\mathbf{b}}\right\rangle_{\widehat{\Pi}},\\ c_0 &= \langle\hat{c}\rangle_{\widehat{\Pi}}.\end{aligned} \tag{3.26}$$

3) Compute the solution $u_0 \in H^1(\Omega)$ of the homogenized boundary value problem

$$-\operatorname{div}\left(\mathbf{A}_0\nabla u_0\right) + \langle B_0, \nabla u_0\rangle + c_0 u_0 = f \quad \text{in } \Omega, \tag{3.27}$$

with $u_0 = g$ on Γ.

4) Compute the **approximation** u_ε^1 of u_ε, which is defined by

$$u_\varepsilon^1(\mathbf{x}) := u_0(\mathbf{x}) - \varepsilon\sum_{j=1}^{d}\widehat{N}_j\left(\frac{\mathbf{x}-\mathbf{x_i}}{\varepsilon}\right)\partial_{x_j}u_0(\mathbf{x}), \quad \forall\mathbf{x}\in\Pi_{\mathbf{i}}^{\varepsilon}, \forall\mathbf{i}. \tag{3.28}$$

The approximation u_ε^1 defined by (3.28) does not satisfy the Dirichlet boundary condition. Therefore, as in [22, p. 28], we correct the approximation with a cutoff function $\varphi_\varepsilon(\mathbf{x})$ satisfying the following conditions:

$$\begin{aligned}&\varphi_\varepsilon \in W_0^{1,\infty}(\Omega), \quad \varphi_\varepsilon \equiv 1 \text{ in } \left\{\mathbf{x}\in\Omega \mid \operatorname{dist}(\mathbf{x},\partial\Omega) > \varepsilon\right\},\\ &\quad 0\le\varphi_\varepsilon\le 1, \quad \varepsilon|\nabla\varphi_\varepsilon| \le c \text{ in } \Omega, \text{ where the constant } c \text{ does not depend on } \varepsilon.\end{aligned} \tag{3.29}$$

Since we consider a domain with Lipschitz boundary, we can take for instance

$$\varphi_\varepsilon(\mathbf{x}) := \min\left\{1, \varepsilon^{-1}\operatorname{dist}(\mathbf{x},\partial\Omega)\right\}. \tag{3.30}$$

Now, the approximation

$$w_\varepsilon^1(\mathbf{x}) := u_0(\mathbf{x}) - \varepsilon\varphi_\varepsilon(\mathbf{x})\left\langle\widehat{\mathbf{N}}\left(\frac{\mathbf{x}-\mathbf{x_i}}{\varepsilon}\right), \nabla u_0(\mathbf{x})\right\rangle, \quad \forall\mathbf{x}\in\Pi_{\mathbf{i}}^{\varepsilon}, \forall\mathbf{i}, \tag{3.31}$$

fulfils the boundary condition.

3.3 A Priori Error Estimate

Theorem 3.1. *Let $f \in L^2(\Omega)$ and u_ε be the solution of (3.4) with $\mathbf{A}_\varepsilon$, $\mathbf{b}_\varepsilon$ and c_ε defined as in Section 3.1. Let u_ε^1 be the asymptotic expansion defined by (3.28) and w_ε^1 be defined by (3.31), where u_0 is the solution of (3.27), $\widehat{N}_j$, for $j = 1, \dots, d$, are the solutions of (3.25) and $\mathbf{A}_0$, B_0 and c_0 are defined by (3.26). Moreover, assume that the derivatives of u_0 up to the third order are in $L^\infty(\Omega)$ and that $\widehat{N}_j \in W^{1,\infty}\left(\widehat{\Pi}\right)$, for $j = 1, \dots, d$. Then, there exist constants c_1, c_2 and c_3 independent of ε such that*

$$\left\|u_\varepsilon - u_\varepsilon^1\right\|_{H^1(\Omega)} \leq c_1 \varepsilon^{\frac{1}{2}}, \quad \left\|u_\varepsilon - w_\varepsilon^1\right\|_{H^1(\Omega)} \leq c_2 \varepsilon^{\frac{1}{2}} \tag{3.32}$$

and

$$\left\|\mathbf{A}_0 \nabla u_0 - \mathbf{A}_\varepsilon \nabla w_\varepsilon^1\right\|_{L^2(\Omega)} \leq c_3 \varepsilon^{\frac{1}{2}}. \tag{3.33}$$

Proof. The proof of both estimates in (3.32) can be found in [22, pp. 26–28], but only for $\mathbf{b}_\varepsilon = \mathbf{0}$ and $c_\varepsilon = 0$.
The first error estimate in (3.32) is also proven in [13, pp. 133–137], under even stronger regularity assumptions. The proof again shows the estimates only for $\mathbf{b}_\varepsilon = \mathbf{0}$ and $c_\varepsilon = 0$. We will show in the following that the arguments used in [13] are still applicable, where we will restrict to the case of homogeneous boundary conditions, i.e., $u_0 \in H_0^1(\Omega)$.
Let us introduce

$$Z_\varepsilon(\mathbf{x}) = u_\varepsilon(\mathbf{x}) - (u_0 + \varepsilon u_1)\left(\mathbf{x}, \frac{\mathbf{x}}{\varepsilon}\right), \tag{3.34}$$

where

$$u_1(\mathbf{x}, \mathbf{y}) = -\sum_{l=1}^{d} \widehat{N}_l(\mathbf{y}) \partial_{x_l} u_0(\mathbf{x}). \tag{3.35}$$

Recall the definition (3.6) of the operators $\mathcal{A}_\varepsilon$ and $\mathcal{B}_\varepsilon$. With the equations (3.7) and (3.8) we calculate $\mathcal{A}_\varepsilon Z_\varepsilon + \mathcal{B}_\varepsilon Z_\varepsilon$:

$$\begin{aligned}
\mathcal{A}_\varepsilon Z_\varepsilon + \mathcal{B}_\varepsilon Z_\varepsilon &= \left(\mathcal{A}_2 + \varepsilon^{-1}\mathcal{A}_1 + \varepsilon^{-2}\mathcal{A}_0\right) Z_\varepsilon + \left(\mathcal{B}_1 + \varepsilon^{-1}\mathcal{B}_0\right) Z_\varepsilon \\
&= \mathcal{A}_\varepsilon u_\varepsilon - \varepsilon^{-2}\mathcal{A}_0 u_0 - \varepsilon^{-1}\left(\mathcal{A}_0 u_1 + \mathcal{A}_1 u_0 + \mathcal{B} u_0\right) \\
&\quad - \left(\mathcal{A}_1 u_1 + \mathcal{A}_2 u_0 + \mathcal{B}_1 u_0 + \mathcal{B}_0 u_1\right) - \varepsilon\left(\mathcal{A}_2 u_1 + \mathcal{B}_1 u_1\right).
\end{aligned}$$

Using (3.11), (3.12) and (3.13), we get

$$\mathcal{A}_\varepsilon Z_\varepsilon(\mathbf{x}) + \mathcal{B}_\varepsilon Z_\varepsilon(\mathbf{x}) = -\varepsilon\left(\mathcal{A}_2 u_1 + \mathcal{B}_1 u_1\right)\left(\mathbf{x}, \frac{\mathbf{x}}{\varepsilon}\right). \tag{3.36}$$

Recall the definition (3.9) of $\mathcal{A}_2$ and (3.10) of $\mathcal{B}_1$. With (3.35) we have:

$$\begin{aligned}
\mathcal{A}_2 u_1 &= \sum_{i,j,l=1}^{d} \hat{a}_{i,j}(\mathbf{y}) \widehat{N}_l(\mathbf{y}) \partial_{x_i} \partial_{x_j} \partial_{x_l} u_0(\mathbf{x}) - \hat{c}(\mathbf{y}) \sum_{l=1}^{d} \widehat{N}_l(\mathbf{y}) \partial_{x_l} u_0(\mathbf{x}), \\
\mathcal{B}_1 u_1 &= -\sum_{i,l=1}^{d} \hat{b}_i(\mathbf{y}) \widehat{N}_l(\mathbf{y}) \partial_{x_i} \partial_{x_l} u_0(\mathbf{x}).
\end{aligned}$$

Due to our regularity assumptions, we have that all the derivatives of u_0 in the equation above are in $L^\infty(\Omega)$.
From (3.36) and since u_0 and u_ε vanish on the boundary $\partial\Omega$, we get the following boundary value problem for Z_ε:

$$\begin{cases} \mathcal{A}_\varepsilon Z_\varepsilon + \mathcal{B}_\varepsilon Z_\varepsilon = \varepsilon F_\varepsilon & \text{in } \Omega, \\ Z_\varepsilon = \varepsilon G_\varepsilon & \text{on } \partial\Omega, \end{cases} \tag{3.37}$$

where

$$F_\varepsilon(\mathbf{x}) = -\sum_{i,j,l=1}^{d} \hat{a}_{i,j}\left(\frac{\mathbf{x}}{\varepsilon}\right)\widehat{N}_l\left(\frac{\mathbf{x}}{\varepsilon}\right)\partial_{x_i}\partial_{x_j}\partial_{x_l}u_0(\mathbf{x}) + \sum_{i,l=1}^{d}\hat{b}_i\left(\frac{\mathbf{x}}{\varepsilon}\right)\widehat{N}_l\left(\frac{\mathbf{x}}{\varepsilon}\right)\partial_{x_i}\partial_{x_l}u_0(\mathbf{x}) + \hat{c}\left(\frac{\mathbf{x}}{\varepsilon}\right)\sum_{l=1}^{d}\widehat{N}_l\left(\frac{\mathbf{x}}{\varepsilon}\right)\partial_{x_l}u_0(\mathbf{x}),$$

$$G_\varepsilon(\mathbf{x}) = \sum_{l=1}^{d}\widehat{N}_l\left(\frac{\mathbf{x}}{\varepsilon}\right)\partial_{x_l}u_0(\mathbf{x}).$$

This is a Dirichlet problem with inhomogeneous boundary condition. Since we have that $\widehat{\mathbf{A}} \in L^\infty\left(\widehat{\Pi}\right)$, $\widehat{N}_j \in W_{\text{per}}\left(\widehat{\Pi}\right)$ for $j = 1, \dots, d$ and that all the derivatives of u_0 are in $L^\infty(\Omega)$, we immediately know that $F_\varepsilon \in L^2(\Omega)$, so we obtain $F_\varepsilon \in H^{-1}(\Omega)$. Moreover, it follows in the same manner as in [13], that $\|F_\varepsilon\|_{H^{-1}(\Omega)} \le C_1$.
In [13], a slightly different problem is considered and thus, the $\widetilde{G}_\varepsilon$ used there is defined as G_ε plus an additional term with second order derivatives of u_0. It is shown that $\widetilde{G}_\varepsilon \in H^{1/2}(\partial\Omega)$ and $\left\|\widetilde{G}_\varepsilon\right\|_{H^{1/2}(\partial\Omega)} \le C_2\varepsilon^{-1/2}$, under the assumption that $\widehat{N}_j \in W^{1,\infty}\left(\widehat{\Pi}\right)$, for $j = 1, \dots, d$. The proof is still valid for G_ε, hence it follows $G_\varepsilon \in H^{1/2}(\partial\Omega)$ and that $\|G_\varepsilon\|_{H^{1/2}(\partial\Omega)} \le C_2\varepsilon^{-1/2}$.
Finally, the following estimate holds for inhomogeneous Dirichlet problems:

$$\begin{aligned}\|Z_\varepsilon\|_{H^1(\Omega)} &\le \varepsilon c_1 \|F_\varepsilon\|_{H^{-1}(\Omega)} + c_2\varepsilon \|G_\varepsilon\|_{H^{1/2}(\partial\Omega)}\\ &\le \varepsilon c_1 C_1 + \varepsilon^{1/2} c_4 C_4\\ &\le c\varepsilon^{1/2},\end{aligned}$$

with a constant c independent of ε, which concludes the proof of (3.32).
Now, we prove (3.33): From [22, p. 27] we know:

$$\left\|\mathbf{A}_0\nabla u_0 - \mathbf{A}_\varepsilon\nabla u_\varepsilon^1\right\|_{L^2(\Omega)} \le c\varepsilon,$$

where c depends on u_0 and

$$\left\|w_\varepsilon^1 - u_\varepsilon^1\right\|_{H^1(\Omega)} \le \frac{c}{\alpha_\varepsilon^{\text{ell}}}\varepsilon^{1/2}.$$

Therefore, we have:

$$\begin{aligned}\left\|\mathbf{A}_0\nabla u_0 - \mathbf{A}_\varepsilon\nabla w_\varepsilon^1\right\|_{L^2(\Omega)} &\le \left\|\mathbf{A}_0\nabla u_0 - \mathbf{A}_\varepsilon\nabla u_\varepsilon^1\right\|_{L^2(\Omega)} + \left\|\mathbf{A}_\varepsilon\nabla\left(u_\varepsilon^1 - w_\varepsilon^1\right)\right\|_{L^2(\Omega)}\\ &\le c\varepsilon + \alpha_\varepsilon^{\text{cont}}\left\|u_\varepsilon^1 - w_\varepsilon^1\right\|_{H^1(\Omega)}\\ &\le c\varepsilon + c\frac{\alpha_\varepsilon^{\text{cont}}}{\alpha_\varepsilon^{\text{ell}}}\varepsilon^{1/2},\end{aligned}$$

which concludes the proof of (3.33). □

Remark 3.2. *The regularity assumptions of Theorem 3.1 are rather strong and we therefore comment on some sufficient conditions.*

1) *For a convex, bounded domain Ω and $f \in L^2(\Omega)$, we know from Theorem 2.33 that the homogeneous Dirichlet problem is H^2-regular, hence, we have $u_0 \in H_0^1(\Omega) \cap H^2(\Omega)$. The much stronger regularity assumption $u_0 \in W^{3,\infty}(\Omega)$ as in Theorem 3.1 can only be guaranteed for much smoother data.*

   In [8, Section 1.5], the a priori error estimate is shown under weaker regularity assumptions, i.e., for $u_0 \in W^{1,\infty}(\Omega)$ instead of $u_0 \in W^{3,\infty}(\Omega)$.

2) *The assumption $\widehat{N}_j \in W^{1,\infty}\left(\widehat{\Pi}\right)$, for $j = 1, \dots, d$, can be deduced with interior regularity estimates from [19], under regularity assumptions on $\widehat{\mathbf{A}}$.*

The properties $u_0 \in H_0^1(\Omega) \cap H^2(\Omega)$ and $\widehat{N}_j \in W_{\text{per}}\left(\widehat{\Pi}\right)$, for $j = 1, \dots, d$, are enough to guarantee u_ε^1 and $w_\varepsilon^1 \in H^1(\Omega)$.

3.4 Properties of the Homogenized Coefficients

To show the subsequent properties, we follow [13, pp. 115–119], while in [8] the proofs are quite similar.

Proposition 3.3. *It holds*

$$(\mathbf{A}_0)_{i,k} = \frac{1}{|\widehat{\Pi}|} \sum_{l,j=1}^{d} \int_{\widehat{\Pi}} \hat{a}_{l,j}(\mathbf{y}) \partial_{y_j} \left(\widehat{N}_k(\mathbf{y}) - y_k \right) \partial_{y_l} \left(\widehat{N}_i(\mathbf{y}) - y_i \right) \mathrm{d}\mathbf{y}, \quad \forall i,k = 1,\dots,d. \tag{3.38}$$

Proof. From the weak formulation of (3.19) we know that $\widehat{N}_j$ is a solution of

$$\int_{\hat{\Pi}} \left\langle \widehat{\mathbf{A}} \nabla_{\mathbf{y}} \widehat{N}_j, \nabla_{\mathbf{y}} v \right\rangle \mathrm{d}\mathbf{y} = \sum_{i=1}^{d} \int_{\hat{\Pi}} \hat{a}_{i,j} \partial_{y_i} v \, \mathrm{d}\mathbf{y}, \quad \forall v \in W_{\mathrm{per}}\left(\widehat{\Pi}\right).$$

Taking as test function $v = \widehat{N}_l$, it follows

$$\int_{\hat{\Pi}} \left\langle \widehat{\mathbf{A}} \nabla_{\mathbf{y}} \widehat{N}_j, \nabla_{\mathbf{y}} \widehat{N}_l \right\rangle \mathrm{d}\mathbf{y} = \sum_{i=1}^{d} \int_{\hat{\Pi}} \hat{a}_{i,j} \partial_{y_i} \widehat{N}_l \, \mathrm{d}\mathbf{y}.$$

The right-hand side can be written equivalently as

$$\sum_{i=1}^{d} \int_{\hat{\Pi}} \hat{a}_{i,j} \partial_{y_i} \widehat{N}_l \, \mathrm{d}\mathbf{y} = \sum_{i,k=1}^{d} \int_{\hat{\Pi}} \hat{a}_{i,k} \partial_{y_k} y_j \partial_{y_i} \widehat{N}_l \, \mathrm{d}\mathbf{y}.$$

Therefore, it follows

$$\sum_{i,k=1}^{d} \int_{\hat{\Pi}} \hat{a}_{i,k} \partial_{y_k} \left(\widehat{N}_j - y_j \right) \partial_{y_i} \widehat{N}_l \, \mathrm{d}\mathbf{y} = 0. \tag{3.39}$$

The definition of $\mathbf{A}_0$ can be rewritten as

$$\begin{aligned}
(\mathbf{A}_0)_{l,j} &= \frac{1}{|\widehat{\Pi}|} \int_{\hat{\Pi}} \left(\hat{a}_{l,j} - \sum_{k=1}^{d} \hat{a}_{l,k} \partial_{y_k} \widehat{N}_j \right) \mathrm{d}\mathbf{y} \\
&= \frac{1}{|\widehat{\Pi}|} \int_{\hat{\Pi}} \left(\sum_{i,k=1}^{d} \hat{a}_{i,k} \partial_{y_k} y_j \partial_{y_i} y_l - \sum_{i,k=1}^{d} \hat{a}_{i,k} \partial_{y_k} \widehat{N}_j \partial_{y_i} y_l \right) \mathrm{d}\mathbf{y} \\
&= \frac{1}{|\widehat{\Pi}|} \sum_{i,k=1}^{d} \int_{\hat{\Pi}} \hat{a}_{i,k} \left(\partial_{y_k} y_j - \partial_{y_k} \widehat{N}_j \right) \partial_{y_i} y_l \, \mathrm{d}\mathbf{y}.
\end{aligned}$$

With equation (3.39) we arrive at

$$(\mathbf{A}_0)_{l,j} = \frac{1}{|\widehat{\Pi}|} \sum_{i,k=1}^{d} \int_{\hat{\Pi}} \hat{a}_{i,k} \partial_{y_k} \left(y_j - \widehat{N}_j \right) \partial_{y_i} \left(y_l - \widehat{N}_l \right) \mathrm{d}\mathbf{y},$$

which completes the proof. □

Proposition 3.4. *The homogenized matrix $\mathbf{A}_0$ is elliptic, i.e., there exists $\alpha_0^{\mathrm{ell}} > 0$ such that*

$$\sum_{i,k=1}^{d} (\mathbf{A}_0)_{i,k} \xi_i \xi_k \geq \alpha_0^{\mathrm{ell}} \|\boldsymbol{\xi}\|_2^2, \quad \forall \boldsymbol{\xi} \in \mathbb{R}^d.$$

Proof. With formula (3.38) it follows:

$$\sum_{i,k=1}^{d} (\mathbf{A}_0)_{i,k} \xi_i \xi_k = \frac{1}{|\widehat{\Pi}|} \sum_{i,k=1}^{d} \sum_{l,j=1}^{d} \int_{\widehat{\Pi}} \hat{a}_{l,j}(\mathbf{y}) \xi_k \partial_{y_j} \left(\widehat{N}_k(\mathbf{y}) - y_k \right) \xi_i \partial_{y_l} \left(\widehat{N}_i(\mathbf{y}) - y_i \right) \mathrm{d}\mathbf{y}.$$

We set $\zeta := \sum_{n=1}^{d} \xi_n \left(\widehat{N}_n(\mathbf{y}) - y_n\right)$ and use the ellipticity of $\widehat{\mathbf{A}}$, hence

$$\begin{aligned}\sum_{i,k=1}^{d} (\mathbf{A}_0)_{i,k}\,\xi_i\xi_k &= \frac{1}{|\widehat{\Pi}|}\sum_{l,j=1}^{d}\int_{\widehat{\Pi}} \hat{a}_{l,j}(\mathbf{y})\partial_{y_j}\zeta\partial_{y_l}\zeta\,\mathrm{d}\mathbf{y}\\ &\geq \frac{\widehat{\alpha}^{\mathrm{ell}}}{|\widehat{\Pi}|}\int_{\widehat{\Pi}}\sum_{l=1}^{d}|\partial_{y_l}\zeta|^2\,\mathrm{d}\mathbf{y}\\ &\geq 0, \quad \forall \boldsymbol{\xi}\in\mathbb{R}^d.\end{aligned}$$

We study the case when

$$\sum_{i,k=1}^{d} (\mathbf{A}_0)_{i,k}\,\xi_i\xi_k = 0,$$

this can only be true if and only if $\partial_{y_l}\zeta = 0,\ \forall l$. Which means that

$$\zeta = \sum_{n=1}^{d}\xi_n\left(\widehat{N}_n(\mathbf{y}) - y_n\right) = \text{constant}$$

and thus

$$\sum_{n=1}^{d}\xi_n\widehat{N}_n(\mathbf{y}) = \sum_{n=1}^{d}\xi_n y_n + \text{constant}.$$

Since $\widehat{N}_n(\mathbf{y})$ is a periodic function, it follows that $\xi_n = 0\ \forall n$. Therefore $\mathbf{A}_0$ is positive definite. □

Proposition 3.5. *If $\widehat{\mathbf{A}}$ is symmetric, then it follows that $\mathbf{A}_0$ is symmetric.*

Proof. This follows straightforwardly from (3.38):

$$\begin{aligned}(\mathbf{A}_0)_{i,k} &= \frac{1}{|\widehat{\Pi}|}\sum_{l,j=1}^{d}\int_{\widehat{\Pi}}\hat{a}_{l,j}(\mathbf{y})\partial_{y_j}\left(\widehat{N}_k(\mathbf{y}) - y_k\right)\partial_{y_l}\left(\widehat{N}_i(\mathbf{y}) - y_i\right)\mathrm{d}\mathbf{y}\\ &=: \frac{1}{|\widehat{\Pi}|}\sum_{l,j=1}^{d}\int_{\widehat{\Pi}}\hat{a}_{l,j}(\mathbf{y})\partial_{y_j}w_k(\mathbf{y})\partial_{y_l}w_i(\mathbf{y})\,\mathrm{d}\mathbf{y}\\ &= \frac{1}{|\widehat{\Pi}|}\sum_{j,l=1}^{d}\int_{\widehat{\Pi}}\hat{a}_{j,l}(\mathbf{y})\partial_{y_l}w_i(\mathbf{y})\partial_{y_j}w_k(\mathbf{y})\,\mathrm{d}\mathbf{y}\\ &= (\mathbf{A}_0)_{k,i}\,.\end{aligned}$$

□

Remark 3.6. *Since we assumed $\hat{c}(\mathbf{y}) \geq \widehat{\alpha} > 0$ for all $\mathbf{y}\in\widehat{\Pi}$ and $-\frac{1}{2}\operatorname{div}\widehat{\mathbf{b}} + \hat{c} \geq 0$, it follows:*

$$c_0 \geq \widehat{\alpha} > 0 \quad \text{and} \quad -\frac{1}{2}\operatorname{div} B_0 + c_0 \geq 0.$$

Proposition 3.4 and in general [13], do not give an explicit constant α_0^{ell}, but, with the help of [22], we can derive upper bounds for $\left(\alpha_0^{\mathrm{ell}}\right)^{-1}$ and α_0^{cont}, as shown in the next proposition. We only consider the case where $\widehat{\mathbf{A}}$ is symmetric, since this always holds in our considerations.

Proposition 3.7. *If $\widehat{\mathbf{A}}$ is symmetric, then it holds*

$$\frac{1}{\alpha_0^{\mathrm{ell}}} \leq \rho\left(\langle\widehat{\mathbf{A}}^{-1}\rangle_{\widehat{\Pi}}\right) \leq \sqrt{\sum_{i,j}^{d}\frac{1}{|\widehat{\Pi}|}\left\|\left(\widehat{\mathbf{A}}^{-1}\right)_{i,j}\right\|^2_{L^2(\widehat{\Pi})}}$$

and

$$\alpha_0^{\mathrm{cont}} \leq \rho\left(\langle\widehat{\mathbf{A}}\rangle_{\widehat{\Pi}}\right) \leq \sqrt{\sum_{i,j}^{d}\frac{1}{|\widehat{\Pi}|}\left\|\hat{a}_{i,j}\right\|^2_{L^2(\widehat{\Pi})}}.$$

Proof. If $\widehat{\mathbf{A}}$ is symmetric, then [22, Section 1.6] gives us the following two-sided estimate:

$$\left\langle \widehat{\mathbf{A}}^{-1} \right\rangle_{\widehat{\Pi}}^{-1} \leq \mathbf{A}_0 \leq \left\langle \widehat{\mathbf{A}} \right\rangle_{\widehat{\Pi}} . \tag{3.40}$$

The notation $\mathbf{B} \leq \mathbf{A}$ stands for $\mathbf{0} \leq \mathbf{A} - \mathbf{B}$ and means that $\mathbf{A} - \mathbf{B}$ is positive semi-definite. From (3.40) it follows immediately:

$$\rho\left(\mathbf{A}_0\right) \leq \rho\left(\left\langle \widehat{\mathbf{A}} \right\rangle_{\widehat{\Pi}}\right).$$

Moreover, we have that

$$\rho\left(\left\langle \widehat{\mathbf{A}} \right\rangle_{\widehat{\Pi}}\right) \leq \left\| \left\langle \widehat{\mathbf{A}} \right\rangle_{\widehat{\Pi}} \right\|_2 \leq \left\| \left\langle \widehat{\mathbf{A}} \right\rangle_{\widehat{\Pi}} \right\|_F ,$$

from (A.3). With the definition of the Frobenius norm and with Hölder's inequality, we get

$$\left\| \left\langle \widehat{\mathbf{A}} \right\rangle_{\widehat{\Pi}} \right\|_F^2 = \sum_{i,j}^{d} \left| \langle \hat{a}_{i,j} \rangle_{\widehat{\Pi}} \right|^2 \leq \sum_{i,j}^{d} \frac{1}{|\widehat{\Pi}|} \left\| \hat{a}_{i,j} \right\|_{L^2(\widehat{\Pi})}^2 .$$

Thus, we have

$$\rho\left(\left\langle \widehat{\mathbf{A}} \right\rangle_{\widehat{\Pi}}\right) \leq \sqrt{\sum_{i,j}^{d} \frac{1}{|\widehat{\Pi}|} \left\| \hat{a}_{i,j} \right\|_{L^2(\widehat{\Pi})}^2}. \tag{3.41}$$

Further, it follows from (3.40) that

$$\mathbf{A}_0^{-1} \leq \left\langle \widehat{\mathbf{A}}^{-1} \right\rangle_{\widehat{\Pi}} ,$$

since we consider positive definite matrices, and therefore

$$\rho\left(\mathbf{A}_0^{-1}\right) \leq \rho\left(\left\langle \widehat{\mathbf{A}}^{-1} \right\rangle_{\widehat{\Pi}}\right).$$

The inequality (3.41) also holds for the inverse matrix, thus we arrive at

$$\frac{1}{\alpha_0^{\mathrm{ell}}} \leq \rho\left(\left\langle \widehat{\mathbf{A}}^{-1} \right\rangle_{\widehat{\Pi}}\right) \leq \sqrt{\sum_{i,j}^{d} \frac{1}{|\widehat{\Pi}|} \left\| \left(\widehat{\mathbf{A}}^{-1}\right)_{i,j} \right\|_{L^2(\widehat{\Pi})}^2},$$

i.e., we have upper bounds for α_0^{cont} and $\left(\alpha_0^{\mathrm{ell}}\right)^{-1}$. □

4 A Posteriori Error of the Two Scale Approximation

In this chapter we want to establish an a posteriori error estimator for the two scale approximation described in the previous chapter. The fully discrete solution of the homogenization problem is not a Galerkin approximation, this is why we consider functional a posteriori estimates, as introduced, in general, in [26]. Further, they have the advantage that they do not require extra regularity and they do not contain mesh-dependent constants. The goal is to find an error estimator that is reliable and can be computed efficiently.
The a posteriori error estimate of functional type for a general reaction-convection-diffusion problem was introduced in Section 2.4. We will first apply this general estimate to the cell and homogenized problems and then conclude with the total error for the homogenization problem.

4.1 Introduction

We consider the problem discussed in Chapter 3, but for simplicity with homogeneous Dirichlet boundary condition:

$$-\operatorname{div}\left(\mathbf{A}_\varepsilon \nabla u_\varepsilon\right) + \langle \mathbf{b}_\varepsilon, \nabla u_\varepsilon \rangle + c_\varepsilon u_\varepsilon = f \quad \text{in } \Pi_{\mathbf{i}}^\varepsilon, \quad \forall \mathbf{i}, \tag{4.1}$$

with $u_\varepsilon = 0$ on Γ, for $f \in L^2(\Omega)$, where the coefficients fulfil the conditions mentioned in Chapter 3. In order to get a two scale approximation w_ε^1 of u_ε, as explained before, we have to solve additionally the cell and the homogenized problems:

1) Compute the solutions $\widehat{N}_k \in W_{\text{per}}\left(\widehat{\Pi}\right)$, $k = 1, \ldots, d$, of the cell problems

$$\begin{aligned} \operatorname{div}\left(\widehat{\mathbf{A}} \nabla \widehat{N}_k\right) &= \operatorname{div}\left(\widehat{\mathbf{A}}_k\right) \quad \text{in } \widehat{\Pi}, \\ \widehat{N}_k &\quad \widehat{\Pi}\text{ – periodic }, \\ \left\langle \widehat{N}_k \right\rangle_{\widehat{\Pi}} &= 0. \end{aligned} \tag{4.2}$$

2) Compute the homogenized coefficients:

$$\begin{aligned} \mathbf{A}_0 &= \left\langle \widehat{\mathbf{A}} \left(I - \left(\nabla \widehat{\mathbf{N}} \right)^\top \right) \right\rangle_{\widehat{\Pi}}, \\ B_0 &= \left\langle \left(I - \nabla \widehat{\mathbf{N}} \right) \widehat{\mathbf{b}} \right\rangle_{\widehat{\Pi}}, \\ c_0 &= \langle \hat{c} \rangle_{\widehat{\Pi}}. \end{aligned} \tag{4.3}$$

3) Compute the solution $u_0 \in H_0^1(\Omega)$ of the homogenized boundary value problem

$$-\operatorname{div}\left(\mathbf{A}_0 \nabla u_0\right) + \langle B_0, \nabla u_0 \rangle + c_0 u_0 = f \quad \text{in } \Omega. \tag{4.4}$$

with $u_0 = 0$ on Γ.

4) Compute the approximation w_ε^1 of u_ε, which is defined by

$$w_\varepsilon^1(\mathbf{x}) := u_0(\mathbf{x}) - \varepsilon \varphi_\varepsilon(\mathbf{x}) \left\langle \widehat{\mathbf{N}}\left(\frac{\mathbf{x} - \mathbf{x}_\mathbf{i}}{\varepsilon} \right), \nabla u_0(\mathbf{x}) \right\rangle, \quad \forall \mathbf{x} \in \Pi_{\mathbf{i}}^\varepsilon, \forall \mathbf{i}. \tag{4.5}$$

For $f \in L^2(\Omega)$ and Ω a convex domain, we know from Theorem 2.33 that $u_0 \in H^2(\Omega)$.
Since we compute a finite dimensional approximation of our problem, we have to consider several errors which arise due to modelling, discretization and the two scale approximation. For that, we introduce some notation:

1) We denote the finite dimensional approximation of $\widehat{\mathbf{N}}$ by $\widehat{\mathbf{N}}^{(l)}$, where $l \in \mathbb{N}$ is an index and we assume that the approximation gets better as l increases.

2) Thus, the computation of the homogenized coefficients is not exact and depends on the accuracy of the approximation $\widehat{\mathbf{N}}^{(l)}$. Hence we compute only an approximation of $\mathbf{A}_0$ and B_0 defined by

$$\mathbf{A}_{0,l} = \left\langle \widehat{\mathbf{A}} \left(I - \left(\nabla \widehat{\mathbf{N}}^{(l)} \right)^{\top} \right) \right\rangle_{\widehat{\Pi}}, \tag{4.6}$$

$$B_{0,l} = \left\langle \left(I - \nabla \widehat{\mathbf{N}}^{(l)} \right) \widehat{\mathbf{b}} \right\rangle_{\widehat{\Pi}}. \tag{4.7}$$

Proposition 4.1. *Let $\widehat{\mathbf{N}}^{(l)}$ be a Galerkin approximation. Then, the approximated homogenized matrix $\mathbf{A}_{0,l}$ is symmetric and elliptic.*

Proof. The proof is straightforward from Propositions 3.4 and 3.5, since $\widehat{\mathbf{N}}^{(l)}$ is a Galerkin approximation. □

Further, since $\mathbf{A}_{0,l}$ is a constant matrix, we can compute $\left(\alpha_{0,l}^{\text{ell}}\right)^{-1} = \rho\left(\mathbf{A}_{0,l}^{-1}\right)$ and $\alpha_{0,l}^{\text{cont}} = \rho\left(\mathbf{A}_{0,l}\right)$.

Remark 4.2. *From Remark 3.6 it follows directly that $-\frac{1}{2}\operatorname{div} B_{0,l} + c_0 \geq 0$.*

3) With these approximated homogenized coefficients, we obtain an approximate homogenized problem which corresponds to the modelling error:

$$-\operatorname{div}\left(\mathbf{A}_{0,l} \nabla u_0^{(l)}\right) + \left\langle B_{0,l}, \nabla u_0^{(l)} \right\rangle + c_0 u_0^{(l)} = f \quad \text{in } \Omega, \tag{4.8}$$

with $u_0^{(l)} = 0$ on Γ. As before, we know that $u_0^{(l)} \in H^2(\Omega)$.

Second, we denote the finite dimensional approximation of $u_0^{(l)}$ by $u_0^{(l,j)}$, where we have again, for $j \in \mathbb{N}$, a better approximation as j increases.

4) The fully discrete approximation is the following finite-dimensional two scale approximation

$$w_{1,\varepsilon}^{(l,j)}(\mathbf{x}) := u_0^{(l,j)}(\mathbf{x}) - \varepsilon \varphi_\varepsilon(\mathbf{x}) \left\langle \widehat{\mathbf{N}}^{(l)}\left(\frac{\mathbf{x}-\mathbf{x_i}}{\varepsilon}\right), \nabla u_0^{(l,j)}(\mathbf{x}) \right\rangle, \quad \forall \mathbf{x} \in \Pi_{\mathbf{i}}^{\varepsilon}, \forall \mathbf{i}. \tag{4.9}$$

5) Since we are interested in estimating $\left\| \nabla\left(u_\varepsilon - w_{1,\varepsilon}^{(l,j)}\right) \right\|_{\mathbf{A}_\varepsilon}$, we have to ensure that $\nabla w_{1,\varepsilon}^{(l,j)}$ is well defined. This is not the case, since $\nabla\left(\nabla u_0^{(l,j)}\right)$ is not defined, because the corresponding finite element space $u_0^{(l,j)} \in V_j \subset H_0^1(\Omega)$ is not contained in $H^2(\Omega)$. Therefore, we need a suitable smoothing operator $\mathbf{P} : L^2(\Omega) \to H^1(\Omega)$. For this purpose, we employ the Clément interpolation operator $\mathbf{C}_h : L^2(\Omega) \to V_j \subset H^1(\Omega)$, see Section A.4. Thus, in our numerical experiments we compute a modified two scale approximation containing a smoothed version of $\nabla u_0^{(l,j)}$:

$$\widetilde{w}_{1,\varepsilon}^{(l,j)}(\mathbf{x}) := u_0^{(l,j)}(\mathbf{x}) - \varepsilon \varphi_\varepsilon(\mathbf{x}) \left\langle \widehat{\mathbf{N}}^{(l)}\left(\frac{\mathbf{x}-\mathbf{x_i}}{\varepsilon}\right), \mathbf{C}_h \nabla u_0^{(l,j)}(\mathbf{x}) \right\rangle, \quad \forall \mathbf{x} \in \Pi_{\mathbf{i}}^{\varepsilon}, \forall \mathbf{i}. \tag{4.10}$$

Our goal is to find an upper bound for the error of the modified two scale approximation, i.e., for

$$\left\| \nabla\left(u_\varepsilon - \widetilde{w}_{1,\varepsilon}^{(l,j)}\right) \right\|_{\mathbf{A}_\varepsilon}.$$

For simplicity, we will consider from now on $\widehat{\mathbf{b}} = \mathbf{0}$ and $\hat{c} = 0$, thus, $B_0 = B_{0,l} = \mathbf{0}$ and $c_0 = 0$, we will generalize the obtained error estimates in Section 4.5. In the following we will assume that f, Ω, $\widehat{\Pi}$ and the matrices $\widehat{\mathbf{A}}$ and $\mathbf{A}_\varepsilon$ fulfil the assumptions made in Section 3.1. Further, let u_ε be the solution of (4.1), let u_0 be the solution of (4.4) and $\widehat{N}_k$ be the solution of (4.2) for $k = 1, \ldots, d$.
In a first estimation step we get the following upper bound:

Theorem 4.3. *For any $v \in H_0^1(\Omega)$ it holds:*

$$\|\nabla (u_\varepsilon - v)\|_{\mathbf{A}_\varepsilon} \leq \|\mathbf{A}_0 \nabla u_0 - \mathbf{A}_\varepsilon \nabla v\|_{\mathbf{A}_\varepsilon^{-1}} .$$

Proof. For any $v, w \in H_0^1(\Omega)$ it holds

$$\begin{aligned}\int_\Omega \langle \mathbf{A}_\varepsilon \nabla (u_\varepsilon - v), \nabla w\rangle &= \int_\Omega f w - \int_\Omega \langle \mathbf{A}_\varepsilon \nabla v, \nabla w\rangle \\ &= \int_\Omega \left(\operatorname{div}(\mathbf{A}_0 \nabla u_0) + f\right) w + \int_\Omega \langle \mathbf{A}_0 \nabla u_0 - \mathbf{A}_\varepsilon \nabla v, \nabla w\rangle ,\end{aligned}$$

since

$$\int_\Omega \langle \mathbf{A}_0 \nabla u_0, \nabla w\rangle + \int_\Omega \operatorname{div}(\mathbf{A}_0 \nabla u_0)\, w = 0.$$

We set $w = u_\varepsilon - v$ and insert $\mathbf{A}_\varepsilon^{1/2} \mathbf{A}_\varepsilon^{-1/2}$ in the second term. Further, with Hölder's inequality, it follows:

$$\begin{aligned}\int_\Omega \langle \mathbf{A}_\varepsilon \nabla (u_\varepsilon - v), \nabla (u_\varepsilon - v)\rangle &\leq \|\operatorname{div}(\mathbf{A}_0 \nabla u_0) + f\|_{L^2(\Omega)} \|u_\varepsilon - v\|_{L^2(\Omega)} \\ &\quad + \int_\Omega \left\langle \mathbf{A}_\varepsilon^{-1/2} (\mathbf{A}_0 \nabla u_0 - \mathbf{A}_\varepsilon \nabla v), \mathbf{A}_\varepsilon^{1/2} \nabla (u_\varepsilon - v)\right\rangle \\ &\leq C_{F\Omega} \|\operatorname{div}(\mathbf{A}_0 \nabla u_0) + f\|_{L^2(\Omega)} \|\nabla (u_\varepsilon - v)\|_{L^2(\Omega)} \\ &\quad + \left(\int_\Omega \left\langle \mathbf{A}_\varepsilon^{-1/2} (\mathbf{A}_0 \nabla u_0 - \mathbf{A}_\varepsilon \nabla v), \mathbf{A}_\varepsilon^{-1/2} (\mathbf{A}_0 \nabla u_0 - \mathbf{A}_\varepsilon \nabla v)\right\rangle\right)^{1/2} \\ &\quad \left(\int_\Omega \left\langle \mathbf{A}_\varepsilon^{1/2} \nabla (u_\varepsilon - v), \mathbf{A}_\varepsilon^{1/2} \nabla (u_\varepsilon - v)\right\rangle\right)^{1/2} \\ &\leq \frac{C_{F\Omega}}{\sqrt{\alpha_\varepsilon^{\mathrm{ell}}}} \|\operatorname{div}(\mathbf{A}_0 \nabla u_0) + f\|_{L^2(\Omega)} \|\nabla (u_\varepsilon - v)\|_{\mathbf{A}_\varepsilon} \\ &\quad + \|\mathbf{A}_\varepsilon \nabla v - \mathbf{A}_0 \nabla u_0\|_{\mathbf{A}_\varepsilon^{-1}} \|\nabla (u_\varepsilon - v)\|_{\mathbf{A}_\varepsilon} .\end{aligned}$$

Dividing this inequality by the norm $\|\nabla (u_\varepsilon - v)\|_{\mathbf{A}_\varepsilon}$, we get

$$\|\nabla (u_\varepsilon - v)\|_{\mathbf{A}_\varepsilon} \leq \frac{C_{F\Omega}}{\sqrt{\alpha_\varepsilon^{\mathrm{ell}}}} \|\operatorname{div}(\mathbf{A}_0 \nabla u_0) + f\|_{L^2(\Omega)} + \|\mathbf{A}_\varepsilon \nabla v - \mathbf{A}_0 \nabla u_0\|_{\mathbf{A}_\varepsilon^{-1}} .$$

Since the homogenized equation is fulfilled, we finally conclude:

$$\|\nabla (u_\varepsilon - v)\|_{\mathbf{A}_\varepsilon} \leq \|\mathbf{A}_0 \nabla u_0 - \mathbf{A}_\varepsilon \nabla v\|_{\mathbf{A}_\varepsilon^{-1}} .$$

□

We can apply this estimate for $v = \widetilde{w}_{1,\varepsilon}^{(l,j)} \in H_0^1(\Omega)$, but it still contains the unknown function u_0. The next theorem is a first step to overcome this problem:

Theorem 4.4. *It holds:*

$$\begin{aligned}\left\|\nabla \left(u_\varepsilon - \widetilde{w}_{1,\varepsilon}^{(l,j)}\right)\right\|_{\mathbf{A}_\varepsilon} &\leq C_1 \rho (\mathbf{A}_0 - \mathbf{A}_{0,l}) + C_2 \left\|\nabla \left(u_0 - u_0^{(l,j)}\right)\right\|_{\mathbf{A}_0} \\ &\quad + \frac{1}{\sqrt{\alpha_\varepsilon^{\mathrm{ell}}}} \left\|\mathbf{A}_{0,l} \nabla u_0^{(l,j)} - \mathbf{A}_\varepsilon \nabla \widetilde{w}_{1,\varepsilon}^{(l,j)}\right\|_{L^2(\Omega)} ,\end{aligned}$$

with

$$C_1 := \frac{C_{F\Omega}}{\sqrt{\alpha_\varepsilon^{\mathrm{ell}}} \sqrt{\alpha_0^{\mathrm{ell}}}} \|f\|_{L^2(\Omega)} , \qquad C_2 := \frac{\alpha_{0,l}^{\mathrm{cont}}}{\sqrt{\alpha_\varepsilon^{\mathrm{ell}}} \sqrt{\alpha_0^{\mathrm{ell}}}}.$$

Proof. Consider the estimate from Theorem 4.3 for $v = \widetilde{w}_{1,\varepsilon}^{(l,j)} \in H_0^1(\Omega)$ and insert the terms $\mathbf{A}_{0,l}\nabla u_0$ and $\mathbf{A}_{0,l}\nabla u_0^{(l,j)}$:

$$\begin{aligned}\left\|\nabla\left(u_\varepsilon - \widetilde{w}_{1,\varepsilon}^{(l,j)}\right)\right\|_{\mathbf{A}_\varepsilon} &\leq \left\|\mathbf{A}_0\nabla u_0 - \mathbf{A}_\varepsilon\nabla\widetilde{w}_{1,\varepsilon}^{(l,j)}\right\|_{\mathbf{A}_\varepsilon^{-1}} \\ &\leq \left\|\left(\mathbf{A}_0 - \mathbf{A}_{0,l}\right)\nabla u_0\right\|_{\mathbf{A}_\varepsilon^{-1}} + \left\|\mathbf{A}_{0,l}\nabla\left(u_0 - u_0^{(l,j)}\right)\right\|_{\mathbf{A}_\varepsilon^{-1}} \\ &\quad + \left\|\mathbf{A}_{0,l}\nabla u_0^{(l,j)} - \mathbf{A}_\varepsilon\nabla\widetilde{w}_{1,\varepsilon}^{(l,j)}\right\|_{\mathbf{A}_\varepsilon^{-1}}.\end{aligned}$$

In order to have suitable weighted norms, we proceed:

$$\begin{aligned}\left\|\nabla\left(u_\varepsilon - \widetilde{w}_{1,\varepsilon}^{(l,j)}\right)\right\|_{\mathbf{A}_\varepsilon} &\leq \frac{1}{\sqrt{\alpha_\varepsilon^{\mathrm{ell}}}}\rho\left(\mathbf{A}_0 - \mathbf{A}_{0,l}\right)\left\|\nabla u_0\right\|_{L^2(\Omega)} \\ &\quad + \frac{\alpha_{0,l}^{\mathrm{cont}}}{\sqrt{\alpha_\varepsilon^{\mathrm{ell}}}\sqrt{\alpha_0^{\mathrm{ell}}}}\left\|\nabla\left(u_0 - u_0^{(l,j)}\right)\right\|_{\mathbf{A}_0} \\ &\quad + \left\|\mathbf{A}_{0,l}\nabla u_0^{(l,j)} - \mathbf{A}_\varepsilon\nabla\widetilde{w}_{1,\varepsilon}^{(l,j)}\right\|_{\mathbf{A}_\varepsilon^{-1}}.\end{aligned}$$

Further, it holds

$$\left\|\nabla u_0\right\|_{L^2(\Omega)} \leq \frac{C_{F\Omega}}{\sqrt{\alpha_0^{\mathrm{ell}}}}\left\|f\right\|_{L^2(\Omega)},$$

which completes the proof. □

This theorem shows that we have three types of errors contributing to the total error:

a) The approximation error of the homogenized matrix $\rho\left(\mathbf{A}_0 - \mathbf{A}_{0,l}\right)$, which is essentially the discretization error of the cell problems estimated in Section 4.2.

b) The combined modelling/discretization error for the homogenized problem $\left\|\nabla\left(u_0 - u_0^{(l,j)}\right)\right\|_{\mathbf{A}_0}$, which we will estimate in Section 4.3.

c) The error term $\left\|\mathbf{A}_{0,l}\nabla u_0^{(l,j)} - \mathbf{A}_\varepsilon\nabla\widetilde{w}_{1,\varepsilon}^{(l,j)}\right\|_{L^2(\Omega)}$, which is computable and also optimal due to equation (3.33) shown in Theorem 3.1.

In the following, we will specify the majorants for the error terms which assemble the total error and conclude in Subsection 4.4 with the total error majorant.
Further, note that one should develop an error estimation strategy (see, e.g., [27]) to balance the above listed error terms, in order to get the desired accuracy of the approximation in an economical way. This means that, if the total error majorant is bigger than the tolerance, one tries to reduce the dominant error term by improving the corresponding approximation.

4.2 Discretization Error for the Cell Problem

Problems with periodic boundary condition, as the cell problem, can only be solved up to a constant. Furthermore, the discretization is similar to the case of Neumann boundary condition, therefore we follow the approach for problems with Neumann boundary condition explained in [26, pp. 80–81]. We define:

$$\left\{\operatorname{div}\widehat{\mathbf{y}}\right\}_{\widehat{\Pi}} := \operatorname{div}\widehat{\mathbf{y}} - \left\langle\operatorname{div}\widehat{\mathbf{y}}\right\rangle_{\widehat{\Pi}}, \tag{4.11}$$

for all $\widehat{\mathbf{y}} \in H\left(\widehat{\Pi}, \operatorname{div}\right)$.

Proposition 4.5 (Discretization error for the cell problem). *The error of the approximations $\widehat{N}_k^{(l)}$ can be estimated by*

$$\left\|\nabla\left(\widehat{N}_k - \widehat{N}_k^{(l)}\right)\right\|_{\widehat{\mathbf{A}}} \leq \mathcal{M}_{\mathrm{disc}}\left(\widehat{N}_k^{(l)}; \widehat{\mathbf{y}}, \widehat{\beta}\right),$$

where

$$\mathcal{M}^2_{\mathrm{disc}}\left(\widehat{N}_k^{(l)};\widehat{\mathbf{y}},\widehat{\beta}\right) \coloneqq \left(1+\widehat{\beta}\right)\left\|\widehat{\mathbf{y}}-\widehat{\mathbf{A}}\nabla\widehat{N}_k^{(l)}\right\|^2_{\widehat{\mathbf{A}}^{-1}} + \frac{C^2_{P\widehat{\Pi}}}{\widehat{\alpha}^{\mathrm{ell}}}\left(1+\frac{1}{\widehat{\beta}}\right)\left\|\left\{\operatorname{div}\widehat{\mathbf{y}}-\operatorname{div}\left(\widehat{\mathbf{A}}_k\right)\right\}_{\widehat{\Pi}}\right\|^2_{L^2(\widehat{\Pi})}$$

for all $\widehat{\mathbf{y}} \in H\left(\widehat{\Pi},\operatorname{div}\right)$, $\widehat{\beta}>0$ *and* $k=1,2,\ldots,d$.

Proof. For any $w \in W_{\mathrm{per}}\left(\widehat{\Pi}\right)$ and $\widehat{\mathbf{y}} \in H\left(\widehat{\Pi},\operatorname{div}\right)$ it holds:

$$\begin{aligned}\int_{\widehat{\Pi}}\left\langle\widehat{\mathbf{A}}\nabla\left(\widehat{N}_k-\widehat{N}_k^{(l)}\right),\nabla w\right\rangle &= \int_{\widehat{\Pi}}-\operatorname{div}\left(\widehat{\mathbf{A}}_k\right)w-\int_{\widehat{\Pi}}\left\langle\widehat{\mathbf{A}}\nabla\widehat{N}_k^{(l)},\nabla w\right\rangle\\ &= \int_{\widehat{\Pi}}\left(\operatorname{div}\widehat{\mathbf{y}}-\operatorname{div}\left(\widehat{\mathbf{A}}_k\right)\right)w+\int_{\widehat{\Pi}}\left\langle\widehat{\mathbf{y}}-\widehat{\mathbf{A}}\nabla\widehat{N}_k^{(l)},\nabla w\right\rangle. \end{aligned} \tag{4.12}$$

Now, let $w=\widehat{N}_k-\widehat{N}_k^{(l)}$. Since $w\in W_{\mathrm{per}}\left(\widehat{\Pi}\right)$, it holds $\langle w\rangle_{\widehat{\Pi}}=0$, hence the first summand is:

$$\begin{aligned}\int_{\widehat{\Pi}}\left(\operatorname{div}\widehat{\mathbf{y}}-\operatorname{div}\left(\widehat{\mathbf{A}}_k\right)\right)\left(\widehat{N}_k-\widehat{N}_k^{(l)}\right) &= \int_{\widehat{\Pi}}\left(\operatorname{div}\widehat{\mathbf{y}}-\operatorname{div}\left(\widehat{\mathbf{A}}_k\right)-\left\langle\operatorname{div}\widehat{\mathbf{y}}-\operatorname{div}\left(\widehat{\mathbf{A}}_k\right)\right\rangle_{\widehat{\Pi}}\right)\left(\widehat{N}_k-\widehat{N}_k^{(l)}\right)\\ &= \int_{\widehat{\Pi}}\left\{\operatorname{div}\widehat{\mathbf{y}}-\operatorname{div}\left(\widehat{\mathbf{A}}_k\right)\right\}_{\widehat{\Pi}}\left(\widehat{N}_k-\widehat{N}_k^{(l)}\right)\\ &\le \left\|\left\{\operatorname{div}\widehat{\mathbf{y}}-\operatorname{div}\left(\widehat{\mathbf{A}}_k\right)\right\}_{\widehat{\Pi}}\right\|_{L^2(\widehat{\Pi})}C_{P\widehat{\Pi}}\left\|\nabla\left(\widehat{N}_k-\widehat{N}_k^{(l)}\right)\right\|_{L^2(\widehat{\Pi})}, \end{aligned}$$

with the definition of $\left\{\cdot\right\}_{\widehat{\Pi}}$ from before. The second summand on the right-hand side of (4.12), with again $w=\widehat{N}_k-\widehat{N}_k^{(l)}$, is estimated by

$$\begin{aligned}\int_{\widehat{\Pi}}\left\langle\widehat{\mathbf{y}}-\widehat{\mathbf{A}}\nabla\widehat{N}_k^{(l)},\nabla\left(\widehat{N}_k-\widehat{N}_k^{(l)}\right)\right\rangle &\le \left(\int_{\widehat{\Pi}}\left\langle\widehat{\mathbf{A}}^{-1/2}\left(\widehat{\mathbf{y}}-\widehat{\mathbf{A}}\nabla\widehat{N}_k^{(l)}\right),\widehat{\mathbf{A}}^{-1/2}\left(\widehat{\mathbf{y}}-\widehat{\mathbf{A}}\nabla\widehat{N}_k^{(l)}\right)\right\rangle\right)^{1/2}\\ &\quad\left(\int_{\widehat{\Pi}}\left\langle\widehat{\mathbf{A}}^{1/2}\nabla\left(\widehat{N}_k-\widehat{N}_k^{(l)}\right),\widehat{\mathbf{A}}^{1/2}\nabla\left(\widehat{N}_k-\widehat{N}_k^{(l)}\right)\right\rangle\right)^{1/2}\\ &\le \left\|\widehat{\mathbf{y}}-\widehat{\mathbf{A}}\nabla\widehat{N}_k^{(l)}\right\|_{\widehat{\mathbf{A}}^{-1}}\left\|\nabla\left(\widehat{N}_k-\widehat{N}_k^{(l)}\right)\right\|_{\widehat{\mathbf{A}}}. \end{aligned}$$

Together we get

$$\begin{aligned}\left\|\nabla\left(\widehat{N}_k-\widehat{N}_k^{(l)}\right)\right\|^2_{\widehat{\mathbf{A}}} &\le \left\|\left\{\operatorname{div}\widehat{\mathbf{y}}-\operatorname{div}\left(\widehat{\mathbf{A}}_k\right)\right\}_{\widehat{\Pi}}\right\|_{L^2(\widehat{\Pi})}\frac{C_{P\widehat{\Pi}}}{\sqrt{\widehat{\alpha}^{\mathrm{ell}}}}\left\|\nabla\left(\widehat{N}_k-\widehat{N}_k^{(l)}\right)\right\|_{\widehat{\mathbf{A}}}\\ &\quad+\left\|\widehat{\mathbf{y}}-\widehat{\mathbf{A}}\nabla\widehat{N}_k^{(l)}\right\|_{\widehat{\mathbf{A}}^{-1}}\left\|\nabla\left(\widehat{N}_k-\widehat{N}_k^{(l)}\right)\right\|_{\widehat{\mathbf{A}}}. \end{aligned}$$

Thus, it follows

$$\left\|\nabla\left(\widehat{N}_k-\widehat{N}_k^{(l)}\right)\right\|_{\widehat{\mathbf{A}}} \le \frac{C_{P\widehat{\Pi}}}{\sqrt{\widehat{\alpha}^{\mathrm{ell}}}}\left\|\left\{\operatorname{div}\widehat{\mathbf{y}}-\operatorname{div}\left(\widehat{\mathbf{A}}_k\right)\right\}_{\widehat{\Pi}}\right\|_{L^2(\widehat{\Pi})}+\left\|\widehat{\mathbf{y}}-\widehat{\mathbf{A}}\nabla\widehat{N}_k^{(l)}\right\|_{\widehat{\mathbf{A}}^{-1}}$$

and by squaring both sides and using Young's inequality we arrive at the definition of the majorant $\mathcal{M}^2_{\mathrm{disc}}\left(\widehat{N}_k^{(l)};\widehat{\mathbf{y}},\widehat{\beta}\right)$. □

The discretization majorant of the cell problems can now be used to approximate the error due to the approximation of the homogenized matrix:

Proposition 4.6 (Approximation error of the homogenized matrix). *The error of the approximation* $\mathbf{A}_{0,l}$ *can be estimated by*

$$\rho\left(\mathbf{A}_0-\mathbf{A}_{0,l}\right) \le \frac{\sqrt{\widehat{\alpha}^{\mathrm{cont}}}}{|\widehat{\Pi}|^{1/2}}\sqrt{\sum_{k=1}^{d}\mathcal{M}^2_{\mathrm{disc}}\left(\widehat{N}_k^{(l)};\widehat{\mathbf{y}},\widehat{\beta}\right)}.$$

Proof. Starting with the definition, we get a first inequality

$$\begin{aligned}\rho\left(\mathbf{A}_0-\mathbf{A}_{0,l}\right) &= \rho\left(\left\langle\widehat{\mathbf{A}}\left(\nabla\left(\widehat{\mathbf{N}}-\widehat{\mathbf{N}}^{(l)}\right)\right)^{\top}\right\rangle_{\widehat{\Pi}}\right)\\ &\leq \left\|\left\langle\widehat{\mathbf{A}}\left(\nabla\left(\widehat{\mathbf{N}}-\widehat{\mathbf{N}}^{(l)}\right)\right)^{\top}\right\rangle_{\widehat{\Pi}}\right\|_2\\ &\leq \frac{1}{|\widehat{\Pi}|}\int_{\widehat{\Pi}}\left\|\widehat{\mathbf{A}}\left(\nabla\left(\widehat{\mathbf{N}}-\widehat{\mathbf{N}}^{(l)}\right)\right)^{\top}\right\|_2\,\mathrm{d}\mathbf{y}.\end{aligned}$$

Now we know that it holds

$$\|\mathbf{AB}\|_2 \leq \|\mathbf{AB}\|_F = \sqrt{\sum_{k=1}^{d}\|\mathbf{Ab}_k\|_2^2},$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Therefore we can further estimate, using also Hölder's inequality in the second line:

$$\begin{aligned}\rho\left(\mathbf{A}_0-\mathbf{A}_{0,l}\right) &\leq \frac{1}{|\widehat{\Pi}|}\int_{\widehat{\Pi}}\sqrt{\sum_{k=1}^{d}\left\|\widehat{\mathbf{A}}\nabla\left(\widehat{N}_k-\widehat{N}_k^{(l)}\right)\right\|_2^2}\,\mathrm{d}\mathbf{y}\\ &\leq \frac{1}{|\widehat{\Pi}|}|\widehat{\Pi}|^{1/2}\sqrt{\sum_{k=1}^{d}\left\|\widehat{\mathbf{A}}\nabla\left(\widehat{N}_k-\widehat{N}_k^{(l)}\right)\right\|_{L^2(\widehat{\Pi})}^2}\\ &\leq \frac{\sqrt{\widehat{\alpha}^{\mathrm{cont}}}}{|\widehat{\Pi}|^{1/2}}\sqrt{\sum_{k=1}^{d}\left\|\nabla\left(\widehat{N}_k-\widehat{N}_k^{(l)}\right)\right\|_{\widehat{\mathbf{A}}}^2}.\end{aligned}$$

The discretization majorant for the cell problems concludes the proof. □

4.3 Modelling/Discretization Error for the Homogenized Problem

For the error estimation of the homogenized problem, we proceed similar to [27]. We set

$$\Lambda_l := \mathbf{A}_{0,l}^{-1/2}\mathbf{A}_0\mathbf{A}_{0,l}^{-1/2}, \quad \kappa_l^2 := 1+\rho\left(\Lambda_l - I\right) \tag{4.13}$$

and formulate the combined error estimate:

Proposition 4.7 (Combined modelling/discretization error for the homogenized problem). *The error of the approximation $u_0^{(l,j)}$ can be estimated by*

$$\begin{aligned}\left\|\nabla\left(u_0-u_0^{(l,j)}\right)\right\|_{\mathbf{A}_0} &\leq E_{\mathrm{disc}}+E_{\mathrm{mod}}\\ &\leq \kappa_l\mathcal{M}_{\mathrm{disc}}\left(u_0^{(l,j)};\mathbf{y},\beta\right)+\delta_{l,j}\rho\left(\mathbf{A}_{0,l}-\mathbf{A}_0\right),\end{aligned}$$

for all $\mathbf{y}\in H(\Omega,\mathrm{div})$ and $\beta>0$, with

$$\delta_{l,j}^2 := \frac{2}{\alpha_0^{\mathrm{ell}}\,\alpha_{0,l}^{\mathrm{ell}}}\left(\left\|\nabla u_0^{(l,j)}\right\|_{\mathbf{A}_{0,l}}^2+\frac{1}{2}\mathcal{M}_{\mathrm{disc}}^2\left(u_0^{(l,j)};\mathbf{y},\beta\right)\right)$$

and the discretization majorant

$$\mathcal{M}_{\mathrm{disc}}^2\left(u_0^{(l,j)};\mathbf{y},\beta\right) := (1+\beta)\left\|\mathbf{A}_{0,l}\nabla u_0^{(l,j)}-\mathbf{y}\right\|_{\mathbf{A}_{0,l}^{-1}}^2+\frac{C_{F\Omega}^2}{\alpha_{0,l}^{\mathrm{ell}}}\left(1+\frac{1}{\beta}\right)\|\mathrm{div}\,\mathbf{y}+f\|_{L^2(\Omega)}^2.$$

Proof. We start with the triangle inequality

$$\left\|\nabla\left(u_0-u_0^{(l,j)}\right)\right\|_{\mathbf{A}_0} \leq \left\|\nabla\left(u_0^{(l)}-u_0^{(l,j)}\right)\right\|_{\mathbf{A}_0}+\left\|\nabla\left(u_0-u_0^{(l)}\right)\right\|_{\mathbf{A}_0} =: E_{\mathrm{disc}}+E_{\mathrm{mod}}.$$

The discretization error can be estimated as follows:

$$\begin{aligned} E_{\text{disc}}^2 &= \left\| \nabla \left(u_0^{(l)} - u_0^{(l,j)} \right) \right\|_{\mathbf{A}_{0,l}}^2 + \int_\Omega \left\langle \left(\mathbf{A}_0 - \mathbf{A}_{0,l} \right) \nabla \left(u_0^{(l)} - u_0^{(l,j)} \right), \nabla \left(u_0^{(l)} - u_0^{(l,j)} \right) \right\rangle \\ &= \left\| \nabla \left(u_0^{(l)} - u_0^{(l,j)} \right) \right\|_{\mathbf{A}_{0,l}}^2 + \int_\Omega \left\langle \left(\Lambda_l - I \right) \mathbf{A}_{0,l}^{1/2} \nabla \left(u_0^{(l)} - u_0^{(l,j)} \right), \mathbf{A}_{0,l}^{1/2} \nabla \left(u_0^{(l)} - u_0^{(l,j)} \right) \right\rangle \\ &\leq \left(1 + \rho \left(\Lambda_l - I \right) \right) \left\| \nabla \left(u_0^{(l)} - u_0^{(l,j)} \right) \right\|_{\mathbf{A}_{0,l}}^2, \end{aligned}$$

where we used $\mathbf{A}_0 - \mathbf{A}_{0,l} = \mathbf{A}_{0,l}^{1/2} \left(\Lambda_l - I \right) \mathbf{A}_{0,l}^{1/2}$. Now, we can apply Theorem 2.37, which gives us the discretization majorant

$$\left\| \nabla \left(u_0^{(l)} - u_0^{(l,j)} \right) \right\|_{\mathbf{A}_{0,l}} \leq \mathcal{M}_{\text{disc}} \left(u_0^{(l,j)}, \mathbf{y} \right) := \left\| \mathbf{y} - \mathbf{A}_{0,l} \nabla u_0^{(l,j)} \right\|_{\mathbf{A}_{0,l}^{-1}} + \frac{C_{F\Omega}}{\sqrt{\alpha_{0,l}^{\text{ell}}}} \left\| f + \operatorname{div} \mathbf{y} \right\|_{L^2(\Omega)}.$$

With a consequence of Young's inequality, namely equation (A.2), we arrive at the estimate:

$$E_{\text{disc}} \leq \kappa_l \mathcal{M}_{\text{disc}} \left(u_0^{(l,j)}; \mathbf{y}, \beta \right).$$

The modelling error is estimated as follows. It holds for any $v \in H_0^1(\Omega)$

$$\begin{aligned} \int_\Omega \left\langle \mathbf{A}_0 \nabla \left(u_0 - u_0^{(l)} \right), \nabla v \right\rangle &= \int_\Omega f v - \int_\Omega \left\langle \mathbf{A}_0 \nabla u_0^{(l)}, \nabla v \right\rangle \\ &= \int_\Omega \left\langle \mathbf{A}_{0,l} \nabla u_0^{(l)}, \nabla v \right\rangle - \int_\Omega \left\langle \mathbf{A}_0 \nabla u_0^{(l)}, \nabla v \right\rangle. \end{aligned}$$

Hence, for $v = u_0 - u_0^{(l)}$:

$$E_{\text{mod}}^2 = \left\| \nabla \left(u_0 - u_0^{(l)} \right) \right\|_{\mathbf{A}_0}^2 = \int_\Omega \left\langle \left(\mathbf{A}_{0,l} - \mathbf{A}_0 \right) \nabla u_0^{(l)}, \nabla \left(u_0 - u_0^{(l)} \right) \right\rangle.$$

Using Hölder's inequality, we arrive at

$$\begin{aligned} \left\| \nabla \left(u_0 - u_0^{(l)} \right) \right\|_{\mathbf{A}_0}^2 &\leq \left(\int_\Omega \left\langle \left(\mathbf{A}_{0,l} - \mathbf{A}_0 \right) \nabla u_0^{(l)}, \left(\mathbf{A}_{0,l} - \mathbf{A}_0 \right) \nabla u_0^{(l)} \right\rangle \right)^{1/2} \\ &\quad \left(\int_\Omega \left\langle \nabla \left(u_0 - u_0^{(l)} \right), \nabla \left(u_0 - u_0^{(l)} \right) \right\rangle \right)^{1/2} \\ &= \rho \left(\mathbf{A}_{0,l} - \mathbf{A}_0 \right) \left\| \nabla u_0^{(l)} \right\|_{L^2(\Omega)} \left\| \nabla \left(u_0 - u_0^{(l)} \right) \right\|_{L^2(\Omega)}. \end{aligned}$$

Further, we want to have weighted norms:

$$\left\| \nabla \left(u_0 - u_0^{(l)} \right) \right\|_{\mathbf{A}_0} \leq \frac{1}{\sqrt{\alpha_{0,l}^{\text{ell}}} \sqrt{\alpha_0^{\text{ell}}}} \rho \left(\mathbf{A}_{0,l} - \mathbf{A}_0 \right) \left\| \nabla u_0^{(l)} \right\|_{\mathbf{A}_{0,l}}.$$

We can expand the norm on the right by a term equal to zero, due to the Galerkin orthogonality, which leads to:

$$\begin{aligned} \left\| \nabla u_0^{(l)} \right\|_{\mathbf{A}_{0,l}}^2 &= \int_\Omega \left\langle \mathbf{A}_{0,l} \nabla u_0^{(l)}, \nabla u_0^{(l)} \right\rangle + \int_\Omega \left\langle \mathbf{A}_{0,l} \nabla u_0^{(l,j)}, \nabla u_0^{(l,j)} \right\rangle - \int_\Omega \left\langle \mathbf{A}_{0,l} \nabla u_0^{(l)}, \nabla u_0^{(l,j)} \right\rangle \\ &= \int_\Omega \left\langle \mathbf{A}_{0,l} \nabla u_0^{(l)}, \nabla \left(u_0^{(l)} - u_0^{(l,j)} \right) \right\rangle + \left\| \nabla u_0^{(l,j)} \right\|_{\mathbf{A}_{0,l}}^2 \\ &\leq \left\| \nabla u_0^{(l)} \right\|_{\mathbf{A}_{0,l}} \left\| \nabla \left(u_0^{(l)} - u_0^{(l,j)} \right) \right\|_{\mathbf{A}_{0,l}} + \left\| \nabla u_0^{(l,j)} \right\|_{\mathbf{A}_{0,l}}^2. \end{aligned}$$

Applying Young's inequality, we get

$$\left\| \nabla u_0^{(l)} \right\|_{\mathbf{A}_{0,l}}^2 \leq \frac{1}{2} \left\| \nabla u_0^{(l)} \right\|_{\mathbf{A}_{0,l}}^2 + \frac{1}{2} \left\| \nabla \left(u_0^{(l)} - u_0^{(l,j)} \right) \right\|_{\mathbf{A}_{0,l}}^2 + \left\| \nabla u_0^{(l,j)} \right\|_{\mathbf{A}_{0,l}}^2.$$

Hence, it follows

$$\left\|\nabla u_0^{(l)}\right\|^2_{\mathbf{A}_{0,l}} \leq 2\left(\left\|\nabla u_0^{(l,j)}\right\|^2_{\mathbf{A}_{0,l}} + \frac{1}{2}\left\|\nabla\left(u_0^{(l)} - u_0^{(l,j)}\right)\right\|^2_{\mathbf{A}_{0,l}}\right).$$

The second term can again be estimated by the discretization majorant, therefore we finally arrive at

$$E_{\mathrm{mod}} \leq \frac{1}{\sqrt{\alpha_{0,l}^{\mathrm{ell}}}\sqrt{\alpha_0^{\mathrm{ell}}}}\rho\left(\mathbf{A}_{0,l} - \mathbf{A}_0\right)\left(2\left(\left\|\nabla u_0^{(l,j)}\right\|^2_{\mathbf{A}_{0,l}} + \frac{1}{2}\mathcal{M}^2_{\mathrm{disc}}\left(u_0^{(l,j)};\mathbf{y},\beta\right)\right)\right)^{1/2}.$$

□

Remark 4.8. *The constant κ_l in the error term E_{disc} could also be defined by $\frac{\rho(\mathbf{A}_0)}{\alpha_{0,l}^{\mathrm{ell}}}$. From both definitions it is clear that κ_l is of size $O(1)$ and since it is multiplied by $\mathcal{M}_{\mathrm{disc}}\left(u_0^{(l,j)};\mathbf{y},\beta\right)$, which goes to zero for $u_0^{(l,j)}$ tending to $u_0^{(l)}$, an upper bound of $O(1)$ is suitable to have a small bound E_{disc}. We will give an a posteriori estimate for κ_l in the following proposition, since $\rho(\mathbf{A}_0)$ is not known.*

Proposition 4.9. *For the coefficient κ_l we have the following computable upper bound:*

$$\kappa_l^2 \leq 1 + \frac{\sqrt{\widehat{\alpha}^{\mathrm{cont}}}}{\alpha_{0,l}^{\mathrm{ell}}\,|\widehat{\Pi}|^{1/2}}\sqrt{\sum_{k=1}^{d}\mathcal{M}^2_{\mathrm{disc}}\left(\widehat{N}_k^{(l)};\widehat{\mathbf{y}},\widehat{\beta}\right)}$$

Proof. Starting with the definition we get a first inequality

$$\begin{aligned}\kappa_l^2 &= 1 + \rho\left(\mathbf{A}_{0,l}^{-1/2}\mathbf{A}_0\mathbf{A}_{0,l}^{-1/2} - I\right)\\ &= 1 + \rho\left(\mathbf{A}_{0,l}^{-1}\left(\mathbf{A}_0 - \mathbf{A}_{0,l}\right)\right)\\ &\leq 1 + \frac{1}{\alpha_{0,l}^{\mathrm{ell}}}\rho\left(\mathbf{A}_0 - \mathbf{A}_{0,l}\right).\end{aligned}$$

Then, we can use the approximation error estimate from Proposition 4.6 and arrive at:

$$\kappa_l^2 \leq 1 + \frac{1}{\alpha_{0,l}^{\mathrm{ell}}}\frac{\sqrt{\widehat{\alpha}^{\mathrm{cont}}}}{|\widehat{\Pi}|^{1/2}}\sqrt{\sum_{k=1}^{d}\mathcal{M}^2_{\mathrm{disc}}\left(\widehat{N}_k^{(l)};\widehat{\mathbf{y}},\widehat{\beta}\right)}.$$

□

Remark 4.10. *The constant $\delta_{l,j}$ in the error term E_{mod} contains the constant value $\left\|\nabla u_0^{(l,j)}\right\|^2_{\mathbf{A}_{0,l}}$, hence, $\delta_{l,j}$ is of size $O(1)$. It is multiplied by $\rho\left(\mathbf{A}_{0,l} - \mathbf{A}_0\right)$, which goes to zero as l increases and is estimated by the upper bound from Proposition 4.6. Therefore, with an upper bound for $\left(\alpha_0^{\mathrm{ell}}\right)^{-1}$, we have a computable error estimate for E_{mod}.*

4.4 Total Error

To conclude, we insert the approximation error of the homogenized matrix from Proposition 4.6 and the combined modelling/discretization error of the homogenized problem from Proposition 4.7 into the estimate of Theorem 4.4 and get the following total error estimate:

Theorem 4.11 (Total error majorant). *Let $\mathbf{A}_\varepsilon$ be defined by (3.3) and condition (3.1) be satisfied. Further, let $\mathbf{A}_0$ be defined by (4.3) and $\mathbf{A}_{0,l}$ be defined by (4.6). We assume that $f \in L^2(\Omega)$, u_ε is the exact solution of (4.1), $u_0^{(l,j)}$ is an approximation of (4.8) and $\widehat{N}_k^{(l)}$, for $k = 1,\dots,d$, is*

an approximation of (4.2). *Further, we assume that the approximation* $\widetilde{w}_{1,\varepsilon}^{(l,j)}$ *is defined by* (4.10). *Then,*

$$\begin{aligned}\left\|\nabla\left(u_\varepsilon - \widetilde{w}_{1,\varepsilon}^{(l,j)}\right)\right\|_{\mathbf{A}_\varepsilon} &\leq \mathcal{M}_{\mathrm{tot}}\left(\widehat{\mathbf{N}}^{(l)}, \mathbf{A}_{0,l}, u_0^{(l,j)}, \widetilde{w}_{1,\varepsilon}^{(l,j)}\right)\\ &:= \left(C_1 + C_2\delta_{l,j}\right)\frac{\sqrt{\widehat{\alpha}^{\mathrm{cont}}}}{|\widehat{\Pi}|^{1/2}}\sqrt{\sum_{k=1}^{d}\mathcal{M}_{\mathrm{disc}}^2\left(\widehat{N}_k^{(l)};\widehat{\mathbf{y}},\widehat{\beta}\right)}\\ &\quad + C_2\kappa_l\mathcal{M}_{\mathrm{disc}}\left(u_0^{(l,j)};\mathbf{y},\beta\right)\\ &\quad + \frac{1}{\sqrt{\alpha_\varepsilon^{\mathrm{ell}}}}\left\|\mathbf{A}_{0,l}\nabla u_0^{(l,j)} - \mathbf{A}_\varepsilon\nabla\widetilde{w}_{1,\varepsilon}^{(l,j)}\right\|_{L^2(\Omega)},\end{aligned}$$

for all $\mathbf{y} \in H(\Omega,\mathrm{div})$, $\widehat{\mathbf{y}} \in H(\widehat{\Pi},\mathrm{div})$ *and* $\beta,\widehat{\beta} > 0$, *with*

$$\begin{aligned}C_1 &:= \frac{C_{F\Omega}}{\sqrt{\alpha_\varepsilon^{\mathrm{ell}}}\sqrt{\alpha_0^{\mathrm{ell}}}}\left\|f\right\|_{L^2(\Omega)}, \quad C_2 := \frac{\alpha_{0,l}^{\mathrm{cont}}}{\sqrt{\alpha_\varepsilon^{\mathrm{ell}}}\sqrt{\alpha_0^{\mathrm{ell}}}},\\ \delta_{l,j}^2 &:= \frac{2}{\alpha_0^{\mathrm{ell}}\,\alpha_{0,l}^{\mathrm{ell}}}\left(\left\|\nabla u_0^{(l,j)}\right\|_{\mathbf{A}_{0,l}}^2 + \frac{1}{2}\mathcal{M}_{\mathrm{disc}}^2\left(u_0^{(l,j)};\mathbf{y},\beta\right)\right).\end{aligned}$$

Hence, the total error majorant is a combination of the discretization majorant $\mathcal{M}_{\mathrm{disc}}\left(\widehat{N}_k^{(l)}\right)$ of the cell problems from Proposition 4.5 and the discretization majorant of the homogenized problem $\mathcal{M}_{\mathrm{disc}}\left(u_0^{(l,j)}\right)$ from Proposition 4.7 and a third computable term, which measures the error of the two scale approximation.

Remark 4.12 (Computability). *The total error majorant from Theorem 4.11 is fully computable, due to the following arguments. Since the matrix* $\mathbf{A}_\varepsilon$ *and the matrix* $\widehat{\mathbf{A}}$ *are given, we can explicitly compute* $\alpha_\varepsilon^{\mathrm{ell}}$, $\widehat{\alpha}^{\mathrm{ell}}$ *and* $\widehat{\alpha}^{\mathrm{cont}}$. *For the constant matrix* $\mathbf{A}_{0,l}$, *we can directly calculate* $\alpha_{0,l}^{\mathrm{ell}}$ *and* $\alpha_{0,l}^{\mathrm{cont}}$. *The values* $\|f\|_{L^2(\Omega)}$, $\left\|\nabla u_0^{(l,j)}\right\|_{\mathbf{A}_{0,l}}^2$ *and* $|\widehat{\Pi}|$ *can be computed explicitly, either through a priori knowledge or through evaluated finite element approximations. Further, Proposition 3.7 gives us a computable a priori bound for* $\left(\alpha_0^{\mathrm{ell}}\right)^{-1}$ *and Proposition 4.9 gives us a computable a posteriori bound for* κ_l. *Finally, for a convex, polygonal domain, we have a priori bounds for the Poincaré and Friedrichs constant, as explained in Section A.3. Hence the total error majorant is fully computable.*

As already mentioned, we want to develop an error estimation strategy, similar to [27]. Assume that we want to solve the original problem (4.1) for a given accuracy δ. For that, we first compute the approximated solutions of the cell problems $\widehat{N}_k^{(l)}$, $k = 1,\dots,d$, and the according majorant $\mathcal{M}_{\mathrm{disc}}\left(\widehat{N}_k^{(l)};\widehat{\mathbf{y}},\widehat{\beta}\right)$ for $l-1$ refinement steps of the initial finite element space V_{h_0}. If this majorant already exceeds δ, then one should increase l and recompute the approximations and the majorant. Otherwise, compute $\mathbf{A}_{0,l}$, the approximated solution of the homogenized problem $u_0^{(l,j)}$ and the according majorant $\mathcal{M}_{\mathrm{disc}}\left(u_0^{(l,j)};\mathbf{y},\beta\right)$ for $j-1$ refinement steps of the initial finite element space V_{H_0}. In a next step, derive the two scale approximation $\widetilde{w}_{1,\varepsilon}^{(l,j)}$ and the total error majorant $\mathcal{M}_{\mathrm{tot}}\left(\widehat{\mathbf{N}}^{(l)},\mathbf{A}_{0,l},u_0^{(l,j)},\widetilde{w}_{1,\varepsilon}^{(l,j)}\right)$ from Theorem 4.11. If the total error majorant exceeds the tolerance δ, then we check which majorant is dominating. If $\mathcal{M}_{\mathrm{disc}}\left(\widehat{N}_k^{(l)}\right) < \nu\mathcal{M}_{\mathrm{disc}}\left(u_0^{(l,j)}\right)$ for some $\nu \in \mathbb{R}_{>0}$, then we should improve the approximation $u_0^{(l,j)}$ by increasing j. If $\mathcal{M}_{\mathrm{disc}}\left(\widehat{N}_k^{(l)}\right) \geq \nu\mathcal{M}_{\mathrm{disc}}\left(u_0^{(l,j)}\right)$, then we should improve the approximation of $\mathbf{A}_{0,l}$ by increasing l and computing more accurate approximations of $\widehat{N}_k^{(l)}$, $k = 1,\dots,d$. With this strategy listed in Algorithm 1, we get an approximated solution $\widetilde{w}_{1,\varepsilon}^{(l,j)}$ of (4.1) for a given accuracy δ, in an economical way.

Algorithm 1 Homogenization error estimation strategy

Input: $\mathbf{A}_\varepsilon$, $\widehat{\mathbf{A}}$, f, Ω, $\widehat{\Pi}$, $\nu \in \mathbb{R}_{>0}$ and $\delta \in (0,1)$.
Compute $\alpha_\varepsilon^{\mathrm{ell}}$, $\widehat{\alpha}^{\mathrm{ell}}$, $\widehat{\alpha}^{\mathrm{cont}}$, $\|f\|_{L^2(\Omega)}$ and $|\widehat{\Pi}|$.
Derive the a priori bounds for $C_{F\Omega}$, $C_{P\widehat{\Pi}}$ and $\left(\alpha_0^{\mathrm{ell}}\right)^{-1}$.

Compute $\widehat{N}_k^{(l)} \in V_{h_l}$ for $k = 1, \dots, d$ and $\mathcal{M}_{\mathrm{disc}}\left(\widehat{N}_k^{(l)}; \widehat{\mathbf{y}}, \widehat{\beta}\right)$.
if $\mathcal{M}_{\mathrm{disc}}\left(\widehat{N}_k^{(l)}; \widehat{\mathbf{y}}, \widehat{\beta}\right) > \delta$ **then**
  Set $l = l + 1$ and start from the beginning.
end if

Compute $\mathbf{A}_{0,l}$ and $\alpha_{0,l}^{\mathrm{ell}}$, $\alpha_{0,l}^{\mathrm{cont}}$

Compute $u_0^{(l,j)} \in V_{H_j}$ and $\mathcal{M}_{\mathrm{disc}}\left(u_0^{(l,j)}; \mathbf{y}, \beta\right)$.
Compute $\left\|\nabla u_0^{(l,j)}\right\|_{\mathbf{A}_{0,l}}^2$, C_1, C_2, $\delta_{l,j}$ and the a posteriori bound of κ_l.
Compute $\widetilde{w}_{1,\varepsilon}^{(l,j)}$ and $\mathcal{M}_{\mathrm{tot}}\left(\widehat{\mathbf{N}}^{(l)}, \mathbf{A}_{0,l}, u_0^{(l,j)}, \widetilde{w}_{1,\varepsilon}^{(l,j)}\right)$.

while $\mathcal{M}_{\mathrm{tot}} > \delta$ **do**
  if $\mathcal{M}_{\mathrm{disc}}\left(\widehat{N}_k^{(l)}\right) < \nu \mathcal{M}_{\mathrm{disc}}\left(u_0^{(l,j)}\right)$ **then**
    Set $j = j + 1$ and return to compute $u_0^{(l,j)} \in V_{H_j}$.
  else if $\mathcal{M}_{\mathrm{disc}}\left(\widehat{N}_k^{(l)}\right) \geq \nu \mathcal{M}_{\mathrm{disc}}\left(u_0^{(l,j)}\right)$ **then**
    Set $l = l + 1$ and start from the beginning.
  end if
end while
Output: $\mathcal{M}_{\mathrm{tot}}\left(\widehat{\mathbf{N}}^{(l)}, \mathbf{A}_{0,l}, u_0^{(l,j)}, \widetilde{w}_{1,\varepsilon}^{(l,j)}\right)$ (Total error majorant)
$\widetilde{w}_{1,\varepsilon}^{(l,j)}$ (Approximation of the homogenization problem)

4.5 Generalized Estimates

In the sections before we considered $\widehat{\mathbf{b}} = \mathbf{0}$ and $\hat{c} = 0$, we will now generalize the estimates for coefficients $\widehat{\mathbf{b}}$ and $\hat{c}$, for which we assume that the assumptions made in Section 3.1 are fulfilled. Hence, we also consider $B_0, B_{0,l}$ and c_0, defined by (4.3) and (4.7). Theorem 4.3 generalizes to:

Theorem 4.13. *For any $v \in H_0^1(\Omega)$ it holds:*

$$\begin{aligned}\|\nabla(u_\varepsilon - v)\|_{\mathbf{A}_\varepsilon} \leq{}& \frac{1}{\sqrt{\alpha_\varepsilon^{\mathrm{ell}}}} \|\mathbf{A}_0 \nabla u_0 - \mathbf{A}_\varepsilon \nabla v\|_{L^2(\Omega)} \\ &+ \frac{C_{F\Omega}}{\sqrt{\alpha_\varepsilon^{\mathrm{ell}}}} \left(\|\langle B_0, \nabla u_0\rangle - \langle \mathbf{b}_\varepsilon, \nabla v\rangle\|_{L^2(\Omega)} + \|c_0 u_0 - c_\varepsilon v\|_{L^2(\Omega)} \right)\end{aligned}$$

Proof. For any $v, w \in H_0^1(\Omega)$ it holds

$$\begin{aligned}&\int_\Omega \left(\langle \mathbf{A}_\varepsilon \nabla(u_\varepsilon - v), \nabla w\rangle + \langle \mathbf{b}_\varepsilon, \nabla(u_\varepsilon - v)\rangle\, w + c_\varepsilon (u_\varepsilon - v) w \right) \\ &= \int_\Omega \left(f - \langle \mathbf{b}_\varepsilon, \nabla v\rangle - c_\varepsilon v \right) w - \int_\Omega \langle \mathbf{A}_\varepsilon \nabla v, \nabla w\rangle .\end{aligned}$$

Since

$$\int_\Omega \langle \mathbf{A}_0 \nabla u_0, \nabla w\rangle + \int_\Omega \left(\langle B_0, \nabla u_0\rangle + c_0 u_0 \right) w + \int_\Omega \operatorname{div}\left(\mathbf{A}_0 \nabla u_0\right) w - \int_\Omega \left(\langle B_0, \nabla u_0\rangle + c_0 u_0 \right) w = 0,$$

it follows

$$\begin{aligned}
&\int_\Omega \left(\langle \mathbf{A}_\varepsilon \nabla(u_\varepsilon - v), \nabla w\rangle + \langle \mathbf{b}_\varepsilon, \nabla(u_\varepsilon - v)\rangle\, w + c_\varepsilon(u_\varepsilon - v)w\right)\\
&= \int_\Omega \left(f + \operatorname{div}\left(\mathbf{A}_0 \nabla u_0\right) - \langle B_0, \nabla u_0\rangle - c_0 u_0\right) w\\
&\quad + \int_\Omega \langle \mathbf{A}_0 \nabla u_0 - \mathbf{A}_\varepsilon \nabla v, \nabla w\rangle\\
&\quad + \int_\Omega \left(\langle B_0, \nabla u_0\rangle - \langle \mathbf{b}_\varepsilon, \nabla v\rangle + c_0 u_0 - c_\varepsilon v\right) w. \qquad (4.14)
\end{aligned}$$

We set $w = u_\varepsilon - v$, then we have for the left-hand side:

$$\begin{aligned}
LHS &:= \int_\Omega \left(\langle \mathbf{A}_\varepsilon \nabla(u_\varepsilon - v), \nabla\left(u_\varepsilon - v\right)\rangle + \langle \mathbf{b}_\varepsilon, \nabla(u_\varepsilon - v)\rangle\left(u_\varepsilon - v\right) + c_\varepsilon(u_\varepsilon - v)^2\right)\\
&= \left\|\nabla\left(u_\varepsilon - v\right)\right\|_{\mathbf{A}_\varepsilon}^2 + \int_\Omega \left(c_\varepsilon - \frac{1}{2}\operatorname{div}\left(\mathbf{b}_\varepsilon\right)\right)\left(u_\varepsilon - v\right)^2\\
&\geq \left\|\nabla\left(u_\varepsilon - v\right)\right\|_{\mathbf{A}_\varepsilon}^2.
\end{aligned}$$

For the first term on the right-hand side of (4.14), we have:

$$\begin{aligned}
&\int_\Omega \left(f + \operatorname{div}\left(\mathbf{A}_0 \nabla u_0\right) - \langle B_0, \nabla u_0\rangle - c_0 u_0\right)\left(u_\varepsilon - v\right)\\
&\leq \left\|f + \operatorname{div}\left(\mathbf{A}_0 \nabla u_0\right) - \langle B_0, \nabla u_0\rangle - c_0 u_0\right\|_{L^2(\Omega)} \left\|u_\varepsilon - v\right\|_{L^2(\Omega)}.
\end{aligned}$$

Since the homogenized equation is fulfilled, this term is equal to zero. For the other terms on the right-hand side of (4.14), we get:

$$\begin{aligned}
RHS &:= \int_\Omega \langle \mathbf{A}_0 \nabla u_0 - \mathbf{A}_\varepsilon \nabla v, \nabla(u_\varepsilon - v)\rangle + \int_\Omega \left(\langle B_0, \nabla u_0\rangle - \langle \mathbf{b}_\varepsilon, \nabla v\rangle + c_0 u_0 - c_\varepsilon v\right)\left(u_\varepsilon - v\right)\\
&\leq \frac{1}{\sqrt{\alpha_\varepsilon^{\mathrm{ell}}}} \left\|\mathbf{A}_0 \nabla u_0 - \mathbf{A}_\varepsilon \nabla v\right\|_{L^2(\Omega)} \left\|\nabla(u_\varepsilon - v)\right\|_{\mathbf{A}_\varepsilon}\\
&\quad + \frac{C_{F\Omega}}{\sqrt{\alpha_\varepsilon^{\mathrm{ell}}}} \left(\left\|\langle B_0, \nabla u_0\rangle - \langle \mathbf{b}_\varepsilon, \nabla v\rangle\right\|_{L^2(\Omega)} + \left\|c_0 u_0 - c_\varepsilon v\right\|_{L^2(\Omega)}\right) \left\|\nabla(u_\varepsilon - v)\right\|_{\mathbf{A}_\varepsilon}.
\end{aligned}$$

Dividing by the norm $\left\|\nabla(u_\varepsilon - v)\right\|_{\mathbf{A}_\varepsilon}$, we conclude:

$$\begin{aligned}
\left\|\nabla\left(u_\varepsilon - v\right)\right\|_{\mathbf{A}_\varepsilon} &\leq \frac{1}{\sqrt{\alpha_\varepsilon^{\mathrm{ell}}}} \left\|\mathbf{A}_0 \nabla u_0 - \mathbf{A}_\varepsilon \nabla v\right\|_{L^2(\Omega)}\\
&\quad + \frac{C_{F\Omega}}{\sqrt{\alpha_\varepsilon^{\mathrm{ell}}}} \left(\left\|\langle B_0, \nabla u_0\rangle - \langle \mathbf{b}_\varepsilon, \nabla v\rangle\right\|_{L^2(\Omega)} + \left\|c_0 u_0 - c_\varepsilon v\right\|_{L^2(\Omega)}\right).
\end{aligned}$$

□

In order to get a computable upper bound for $\left\|\nabla\left(u_\varepsilon - \widetilde{w}_{1,\varepsilon}^{(l,j)}\right)\right\|_{\mathbf{A}_\varepsilon}$, we proceed similar to Theorem 4.4:

Theorem 4.14. *It holds:*

$$\begin{aligned}
\left\|\nabla\left(u_\varepsilon - \widetilde{w}_{1,\varepsilon}^{(l,j)}\right)\right\|_{\mathbf{A}_\varepsilon} &\leq C_1 \left(\rho\left(\mathbf{A}_0 - \mathbf{A}_{0,l}\right) + C_{F\Omega}\left\|B_0 - B_{0,l}\right\|_{L^\infty(\Omega)}\right) + C_2 \left\|\nabla\left(u_0 - u_0^{(l,j)}\right)\right\|_{\mathbf{A}_0}\\
&\quad + \frac{1}{\sqrt{\alpha_\varepsilon^{\mathrm{ell}}}} \left\|\mathbf{A}_{0,l} \nabla u_0^{(l,j)} - \mathbf{A}_\varepsilon \nabla \widetilde{w}_{1,\varepsilon}^{(l,j)}\right\|_{L^2(\Omega)}\\
&\quad + \frac{C_{F\Omega}}{\sqrt{\alpha_\varepsilon^{\mathrm{ell}}}} \left(\left\|\left\langle B_{0,l}, \nabla u_0^{(l,j)}\right\rangle - \left\langle \mathbf{b}_\varepsilon, \nabla \widetilde{w}_{1,\varepsilon}^{(l,j)}\right\rangle\right\|_{L^2(\Omega)} + \left\|c_0 u_0^{(l,j)} - c_\varepsilon \widetilde{w}_{1,\varepsilon}^{(l,j)}\right\|_{L^2(\Omega)}\right),
\end{aligned}$$

with

$$C_1 := \frac{C_{F\Omega}}{\sqrt{\alpha_\varepsilon^{\mathrm{ell}}}\sqrt{\alpha_0^{\mathrm{ell}}}} \left\|f\right\|_{L^2(\Omega)}, \qquad C_2 := \frac{1}{\sqrt{\alpha_\varepsilon^{\mathrm{ell}}}\sqrt{\alpha_0^{\mathrm{ell}}}} \left(\alpha_{0,l}^{\mathrm{cont}} + C_{F\Omega}\left\|B_{0,l}\right\|_{L^\infty(\Omega)} + C_{F\Omega}^2\left\|c_0\right\|_{L^\infty(\Omega)}\right).$$

Proof. We use the estimate from Theorem 4.13 for $v = \widetilde{w}_{1,\varepsilon}^{(l,j)} \in H_0^1(\Omega)$, insert known values and estimate further:

$$\begin{aligned}
\sqrt{\alpha_\varepsilon^{\mathrm{ell}}} \left\| \nabla \left(u_\varepsilon - \widetilde{w}_{1,\varepsilon}^{(l,j)} \right) \right\|_{\mathbf{A}_\varepsilon} &\leq \left\| \mathbf{A}_0 \nabla u_0 - \mathbf{A}_\varepsilon \nabla \widetilde{w}_{1,\varepsilon}^{(l,j)} \right\|_{L^2(\Omega)} \\
&\quad + C_{F\Omega} \left(\left\| \langle B_0, \nabla u_0 \rangle - \left\langle \mathbf{b}_\varepsilon, \nabla \widetilde{w}_{1,\varepsilon}^{(l,j)} \right\rangle \right\|_{L^2(\Omega)} + \left\| c_0 u_0 - c_\varepsilon \widetilde{w}_{1,\varepsilon}^{(l,j)} \right\|_{L^2(\Omega)} \right) \\
&\leq \left\| \left(\mathbf{A}_0 - \mathbf{A}_{0,l} \right) \nabla u_0 \right\|_{L^2(\Omega)} + \left\| \mathbf{A}_{0,l} \nabla \left(u_0 - u_0^{(l,j)} \right) \right\|_{L^2(\Omega)} \\
&\quad + \left\| \mathbf{A}_{0,l} \nabla u_0^{(l,j)} - \mathbf{A}_\varepsilon \nabla \widetilde{w}_{1,\varepsilon}^{(l,j)} \right\|_{L^2(\Omega)} \\
&\quad + C_{F\Omega} \left(\left\| \langle B_0 - B_{0,l}, \nabla u_0 \rangle \right\|_{L^2(\Omega)} + \left\| \left\langle B_{0,l}, \nabla \left(u_0 - u_0^{(l,j)} \right) \right\rangle \right\|_{L^2(\Omega)} \right) \\
&\quad + C_{F\Omega} \left\| \left\langle B_{0,l}, \nabla u_0^{(l,j)} \right\rangle - \left\langle \mathbf{b}_\varepsilon, \nabla \widetilde{w}_{1,\varepsilon}^{(l,j)} \right\rangle \right\|_{L^2(\Omega)} \\
&\quad + C_{F\Omega} \left(\left\| c_0 \left(u_0 - u_0^{(l,j)} \right) \right\|_{L^2(\Omega)} + \left\| c_0 u_0^{(l,j)} - c_\varepsilon \widetilde{w}_{1,\varepsilon}^{(l,j)} \right\|_{L^2(\Omega)} \right).
\end{aligned}$$

Since it holds

$$\left\| \langle \mathbf{v}, \mathbf{w} \rangle \right\|_{L^2(\Omega)} \leq \left\| \mathbf{v} \right\|_{L^{2p}(\Omega)} \left\| \mathbf{w} \right\|_{L^{2q}(\Omega)},$$

for $1 \leq p \leq \infty$ and q its conjugate exponent (see Definition A.3), we get:

$$\begin{aligned}
\sqrt{\alpha_\varepsilon^{\mathrm{ell}}} \left\| \nabla \left(u_\varepsilon - \widetilde{w}_{1,\varepsilon}^{(l,j)} \right) \right\|_{\mathbf{A}_\varepsilon} &\leq \left(\rho \left(\mathbf{A}_0 - \mathbf{A}_{0,l} \right) + C_{F\Omega} \left\| B_0 - B_{0,l} \right\|_{L^\infty(\Omega)} \right) \left\| \nabla u_0 \right\|_{L^2(\Omega)} \\
&\quad + \frac{1}{\sqrt{\alpha_0^{\mathrm{ell}}}} \left(\alpha_{0,l}^{\mathrm{cont}} + C_{F\Omega} \left\| B_{0,l} \right\|_{L^\infty(\Omega)} + C_{F\Omega}^2 \left\| c_0 \right\|_{L^\infty(\Omega)} \right) \left\| \nabla \left(u_0 - u_0^{(l,j)} \right) \right\|_{\mathbf{A}_0} \\
&\quad + \left\| \mathbf{A}_{0,l} \nabla u_0^{(l,j)} - \mathbf{A}_\varepsilon \nabla \widetilde{w}_{1,\varepsilon}^{(l,j)} \right\|_{L^2(\Omega)} \\
&\quad + C_{F\Omega} \left\| \left\langle B_{0,l}, \nabla u_0^{(l,j)} \right\rangle - \left\langle \mathbf{b}_\varepsilon, \nabla \widetilde{w}_{1,\varepsilon}^{(l,j)} \right\rangle \right\|_{L^2(\Omega)} \\
&\quad + C_{F\Omega} \left\| c_0 u_0^{(l,j)} - c_\varepsilon \widetilde{w}_{1,\varepsilon}^{(l,j)} \right\|_{L^2(\Omega)}.
\end{aligned}$$

As before, we use the a priori bound

$$\left\| \nabla u_0 \right\|_{L^2(\Omega)} \leq \frac{C_{F\Omega}}{\sqrt{\alpha_0^{\mathrm{ell}}}} \left\| f \right\|_{L^2(\Omega)},$$

which concludes the proof. □

This theorem shows that we have some additional types of error terms and some further changes, as summarized in the following:

a) Derive an approximation error of the homogenized coefficient $\left\| B_0 - B_{0,l} \right\|_{L^\infty(\Omega)}$.

b) Derive a new combined modelling/discretization error for the generalized homogenized problem $\left\| \nabla \left(u_0 - u_0^{(l,j)} \right) \right\|_{\mathbf{A}_0}$.

c) The error terms $\left\| \left\langle B_{0,l}, \nabla u_0^{(l,j)} \right\rangle - \left\langle \mathbf{b}_\varepsilon, \nabla \widetilde{w}_{1,\varepsilon}^{(l,j)} \right\rangle \right\|_{L^2(\Omega)}$ and $\left\| c_0 u_0^{(l,j)} - c_\varepsilon \widetilde{w}_{1,\varepsilon}^{(l,j)} \right\|_{L^2(\Omega)}$ are computable.

Below, we will specify those additional error terms. Since we do not get new unknown constants, we can then directly conclude with the generalized total error majorant.

Proposition 4.15 (Approximation error of the homogenized coefficient). *The error of the approximation $B_{0,l}$ can be approximated by*

$$\left\| B_0 - B_{0,l} \right\|_{L^\infty(\Omega)} \leq \frac{\left\| \mathbf{b}_\varepsilon \right\|_{L^\infty(\Omega)}}{\sqrt{\widehat{\alpha}^{\mathrm{ell}}} \left| \widehat{\Pi} \right|^{1/2}} \max_{1 \leq k \leq d} \left| \mathcal{M}_{\mathrm{disc}} \left(\widehat{N}_k^{(l)}; \widehat{\mathbf{y}}, \widehat{\beta} \right) \right|.$$

Proof. We start with the definition:

$$\begin{aligned}\|B_0 - B_{0,l}\|_{L^\infty(\Omega)} &= \left\|\left\langle \nabla\left(\widehat{\mathbf{N}}^{(l)} - \widehat{\mathbf{N}}\right)\widehat{\mathbf{b}}\right\rangle_{\widehat{\Pi}}\right\|_{L^\infty(\Omega)} \\ &= \max_{1\le k\le d} \operatorname*{ess\,sup}_{\mathbf{x}\in\Omega} \left| \frac{1}{|\widehat{\Pi}|}\int_{\widehat{\Pi}} \left(\nabla\left(\widehat{\mathbf{N}}^{(l)} - \widehat{\mathbf{N}}\right)\widehat{\mathbf{b}}\right)_k \mathrm{d}\mathbf{y}\right| \\ &= \max_{1\le k\le d} \operatorname*{ess\,sup}_{\mathbf{x}\in\Omega} \left| \frac{1}{|\widehat{\Pi}|}\int_{\widehat{\Pi}} \left\langle \nabla\left(\widehat{N}_k^{(l)} - \widehat{N}_k\right), \widehat{\mathbf{b}}\right\rangle \mathrm{d}\mathbf{y}\right| \\ &\le \max_{1\le k\le d} \operatorname*{ess\,sup}_{\mathbf{x}\in\Omega} \left| \frac{1}{|\widehat{\Pi}|}\int_{\widehat{\Pi}} \left\|\nabla\left(\widehat{N}_k^{(l)} - \widehat{N}_k\right)\right\|_2 \left\|\widehat{\mathbf{b}}\right\|_2 \mathrm{d}\mathbf{y}\right|.\end{aligned}$$

Since $\widehat{\mathbf{b}}(\mathbf{y}) = \mathbf{b}_\varepsilon(\mathbf{x})$, it follows:

$$\begin{aligned}\|B_0 - B_{0,l}\|_{L^\infty(\Omega)} &\le \|\mathbf{b}_\varepsilon\|_{L^\infty(\Omega)} \max_{1\le k\le d} \left| \frac{1}{|\widehat{\Pi}|}\int_{\widehat{\Pi}} \left\|\nabla\left(\widehat{N}_k^{(l)} - \widehat{N}_k\right)\right\|_2 \mathrm{d}\mathbf{y}\right| \\ &\le \|\mathbf{b}_\varepsilon\|_{L^\infty(\Omega)} \max_{1\le k\le d} \left| \frac{1}{|\widehat{\Pi}|^{1/2}} \left\|\nabla\left(\widehat{N}_k^{(l)} - \widehat{N}_k\right)\right\|_{L^2(\widehat{\Pi})}\right| \\ &\le \|\mathbf{b}_\varepsilon\|_{L^\infty(\Omega)} \frac{1}{\sqrt{\widehat{\alpha}^{\mathrm{ell}}}\,|\widehat{\Pi}|^{1/2}} \max_{1\le k\le d} \left|\mathcal{M}_{\mathrm{disc}}\left(\widehat{N}_k^{(l)}; \widehat{\mathbf{y}}, \widehat{\beta}\right)\right|.\end{aligned}$$

□

As before, we set

$$\Lambda_l := \mathbf{A}_{0,l}^{-1/2}\mathbf{A}_0\mathbf{A}_{0,l}^{-1/2}, \quad \kappa_l^2 := 1 + \rho\left(\Lambda_l - I\right).$$

Similar to Proposition 4.7, we get the combined error estimate:

Proposition 4.16 (New combined modelling/discretization error for the homogenized problem). *The error of the approximation $u_0^{(l,j)}$ can be estimated by*

$$\begin{aligned}\left\|\nabla\left(u_0 - u_0^{(l,j)}\right)\right\|_{\mathbf{A}_0} &\le E_{\mathrm{disc}} + E_{\mathrm{mod}} \\ &\le \kappa_l \mathcal{M}_{\mathrm{disc}}\left(u_0^{(l,j)}; \mathbf{y}, \beta\right) + \delta_{l,j}\left(\rho\left(\mathbf{A}_{0,l} - \mathbf{A}_0\right) + C_{F\Omega}\left\|B_{0,l} - B_0\right\|_{L^\infty(\Omega)}\right),\end{aligned}$$

for all $\mathbf{y} \in H(\Omega, \mathrm{div})$ and $\beta > 0$, with

$$\delta_{l,j} := \frac{1}{\sqrt{\alpha_0^{\mathrm{ell}}}\sqrt{\alpha_{0,l}^{\mathrm{ell}}}}\left(\mathcal{M}_{\mathrm{disc}}\left(u_0^{(l,j)}; \mathbf{y}, \beta\right) + \left\|\nabla u_0^{(l,j)}\right\|_{\mathbf{A}_{0,l}}\right).$$

The discretization majorant is denoted by

$$\begin{aligned}\mathcal{M}_{\mathrm{disc}}^2\left(u_0^{(l,j)}; \mathbf{y}, \beta\right) := (1+\beta)\left\|\mathbf{A}_{0,l}\nabla u_0^{(l,j)} - \mathbf{y}\right\|_{\mathbf{A}_{0,l}^{-1}}^2 \\ + \frac{C_{F\Omega}^2}{\alpha_{0,l}^{\mathrm{ell}}}\left(1 + \frac{1}{\beta}\right)\left\|\mathrm{div}\,\mathbf{y} - \left\langle B_{0,l}, \nabla u_0^{(l,j)}\right\rangle - c_0 u_0^{(l,j)} + f\right\|_{L^2(\Omega)}^2.\end{aligned}$$

Proof. We start with the triangle inequality

$$\left\|\nabla\left(u_0 - u_0^{(l,j)}\right)\right\|_{\mathbf{A}_0} \le \left\|\nabla\left(u_0^{(l)} - u_0^{(l,j)}\right)\right\|_{\mathbf{A}_0} + \left\|\nabla\left(u_0 - u_0^{(l)}\right)\right\|_{\mathbf{A}_0} =: E_{\mathrm{disc}} + E_{\mathrm{mod}}.$$

The discretization error can be estimated as follows:

$$\begin{aligned}E_{\mathrm{disc}}^2 &= \left\|\nabla\left(u_0^{(l)} - u_0^{(l,j)}\right)\right\|_{\mathbf{A}_{0,l}}^2 + \int_\Omega \left\langle\left(\mathbf{A}_0 - \mathbf{A}_{0,l}\right)\nabla\left(u_0^{(l)} - u_0^{(l,j)}\right), \nabla\left(u_0^{(l)} - u_0^{(l,j)}\right)\right\rangle \\ &= \left\|\nabla\left(u_0^{(l)} - u_0^{(l,j)}\right)\right\|_{\mathbf{A}_{0,l}}^2 + \int_\Omega \left\langle\left(\Lambda_l - I\right)\mathbf{A}_{0,l}^{1/2}\nabla\left(u_0^{(l)} - u_0^{(l,j)}\right), \mathbf{A}_{0,l}^{1/2}\nabla\left(u_0^{(l)} - u_0^{(l,j)}\right)\right\rangle \\ &\le \left(1 + \rho\left(\Lambda_l - I\right)\right)\left\|\nabla\left(u_0^{(l)} - u_0^{(l,j)}\right)\right\|_{\mathbf{A}_{0,l}}^2,\end{aligned}$$

where we used $\mathbf{A}_0 - \mathbf{A}_{0,l} = \mathbf{A}_{0,l}^{1/2}\left(\Lambda_l - I\right)\mathbf{A}_{0,l}^{1/2}$. Now, we can apply Theorem 2.37, which gives us the discretization majorant

$$\left\|\nabla\left(u_0^{(l)} - u_0^{(l,j)}\right)\right\|_{\mathbf{A}_{0,l}} \leq \mathcal{M}_{\mathrm{disc}}\left(u_0^{(l,j)}, \mathbf{y}\right) := \left\|\mathbf{y} - \mathbf{A}_{0,l}\nabla u_0^{(l,j)}\right\|_{\mathbf{A}_{0,l}^{-1}} + \frac{C_{F\Omega}}{\sqrt{\alpha_{0,l}^{\mathrm{ell}}}}\left\|f - \left\langle B_{0,l}, \nabla u_0^{(l,j)}\right\rangle - c_0 u_0^{(l,j)} + \operatorname{div}\mathbf{y}\right\|_{L^2(\Omega)}.$$

With a consequence of Young's inequality, namely equation (A.2), we arrive at the estimate:

$$E_{\mathrm{disc}} \leq \kappa_l \mathcal{M}_{\mathrm{disc}}\left(u_0^{(l,j)}; \mathbf{y}, \beta\right).$$

The modelling error is estimated as follows. It holds for any $v \in H_0^1(\Omega)$:

$$\int_\Omega \left(\langle B_0, \nabla v\rangle v + c_0 v^2\right) = \int_\Omega \left(c_0 - \frac{1}{2}\operatorname{div} B_0\right) v^2 \geq 0.$$

Hence, for $v = u_0 - u_0^{(l)}$:

$$\begin{aligned} E_{\mathrm{mod}}^2 &= \left\|\nabla\left(u_0 - u_0^{(l)}\right)\right\|_{\mathbf{A}_0}^2 \\ &\leq \left\|\nabla\left(u_0 - u_0^{(l)}\right)\right\|_{\mathbf{A}_0}^2 + \int_\Omega \left(\left\langle B_0, \nabla\left(u_0 - u_0^{(l)}\right)\right\rangle\left(u_0 - u_0^{(l)}\right) + c_0\left(u_0 - u_0^{(l)}\right)^2\right). \end{aligned}$$

Further, since u_0 is the exact solution of (4.4) and $u_0^{(l)}$ is the exact solution of (4.8):

$$\begin{aligned} E_{\mathrm{mod}}^2 &\leq \int_\Omega \left(\left\langle \mathbf{A}_0\nabla\left(u_0 - u_0^{(l)}\right), \nabla\left(u_0 - u_0^{(l)}\right)\right\rangle + \left\langle B_0, \nabla\left(u_0 - u_0^{(l)}\right)\right\rangle\left(u_0 - u_0^{(l)}\right) + c_0\left(u_0 - u_0^{(l)}\right)^2\right) \\ &= \int_\Omega \left(f - c_0 u_0^{(l)}\right)\left(u_0 - u_0^{(l)}\right) - \int_\Omega \left\langle \mathbf{A}_0\nabla u_0^{(l)}, \nabla\left(u_0 - u_0^{(l)}\right)\right\rangle - \int_\Omega \left\langle B_0, \nabla u_0^{(l)}\right\rangle\left(u_0 - u_0^{(l)}\right) \\ &= \int_\Omega \left(f - \left\langle B_{0,l}, \nabla u_0^{(l)}\right\rangle - c_0 u_0^{(l)}\right)\left(u_0 - u_0^{(l)}\right) \\ &\quad - \int_\Omega \left\langle \mathbf{A}_0\nabla u_0^{(l)}, \nabla\left(u_0 - u_0^{(l)}\right)\right\rangle + \int_\Omega \left\langle B_{0,l} - B_0, \nabla u_0^{(l)}\right\rangle\left(u_0 - u_0^{(l)}\right) \\ &= \int_\Omega \left\langle \left(\mathbf{A}_{0,l} - \mathbf{A}_0\right)\nabla u_0^{(l)}, \nabla\left(u_0 - u_0^{(l)}\right)\right\rangle + \int_\Omega \left\langle B_{0,l} - B_0, \nabla u_0^{(l)}\right\rangle\left(u_0 - u_0^{(l)}\right). \end{aligned}$$

Using for the first term Hölder's inequality and for the second term again

$$\|\langle \mathbf{v}, \mathbf{w}\rangle\|_{L^2(\Omega)} \leq \|\mathbf{v}\|_{L^{2p}(\Omega)}\|\mathbf{w}\|_{L^{2q}(\Omega)},$$

for $1 \leq p \leq \infty$ and q its conjugate exponent, we arrive at

$$\begin{aligned} \left\|\nabla\left(u_0 - u_0^{(l)}\right)\right\|_{\mathbf{A}_0}^2 &\leq \rho\left(\mathbf{A}_{0,l} - \mathbf{A}_0\right)\left\|\nabla u_0^{(l)}\right\|_{L^2(\Omega)}\left\|\nabla\left(u_0 - u_0^{(l)}\right)\right\|_{L^2(\Omega)} \\ &\quad + \|B_{0,l} - B_0\|_{L^\infty(\Omega)}\left\|\nabla u_0^{(l)}\right\|_{L^2(\Omega)} C_{F\Omega}\left\|\nabla\left(u_0 - u_0^{(l)}\right)\right\|_{L^2(\Omega)}. \end{aligned}$$

Hence,

$$\begin{aligned} \left\|\nabla\left(u_0 - u_0^{(l)}\right)\right\|_{\mathbf{A}_0} &\leq \frac{1}{\sqrt{\alpha_0^{\mathrm{ell}}}}\left(\rho\left(\mathbf{A}_{0,l} - \mathbf{A}_0\right) + C_{F\Omega}\|B_{0,l} - B_0\|_{L^\infty(\Omega)}\right)\left\|\nabla u_0^{(l)}\right\|_{L^2(\Omega)} \\ &\leq \frac{1}{\sqrt{\alpha_0^{\mathrm{ell}}}}\left(\rho\left(\mathbf{A}_{0,l} - \mathbf{A}_0\right) + C_{F\Omega}\|B_{0,l} - B_0\|_{L^\infty(\Omega)}\right)\frac{1}{\sqrt{\alpha_{0,l}^{\mathrm{ell}}}}\left\|\nabla u_0^{(l)}\right\|_{\mathbf{A}_{0,l}}. \end{aligned}$$

Since it holds

$$\left\|\nabla u_0^{(l)}\right\|_{\mathbf{A}_{0,l}} \leq \left\|\nabla\left(u_0^{(l)} - u_0^{(l,j)}\right)\right\|_{\mathbf{A}_{0,l}} + \left\|\nabla u_0^{(l,j)}\right\|_{\mathbf{A}_{0,l}}$$

and

$$\left\|\nabla\left(u_0^{(l)}-u_0^{(l,j)}\right)\right\|_{\mathbf{A}_{0,l}}\leq\mathcal{M}_{\mathrm{disc}}\left(u_0^{(l,j)};\mathbf{y},\beta\right),$$

we finally arrive at

$$E_{\mathrm{mod}}\leq\frac{1}{\sqrt{\alpha_0^{\mathrm{ell}}}\sqrt{\alpha_{0,l}^{\mathrm{ell}}}}\left(\rho\left(\mathbf{A}_{0,l}-\mathbf{A}_0\right)+C_{F\Omega}\left\|B_{0,l}-B_0\right\|_{L^\infty(\Omega)}\right)\left(\mathcal{M}_{\mathrm{disc}}\left(u_0^{(l,j)};\mathbf{y},\beta\right)+\left\|\nabla u_0^{(l,j)}\right\|_{\mathbf{A}_{0,l}}\right).$$

□

Remark 4.17. *The constants κ_l and $\delta_{l,j}$ are defined similar to Proposition 4.7, therefore they are again of size $O(1)$, as mentioned in Remark 4.8 and 4.10. They are multiplied by $\mathcal{M}_{\mathrm{disc}}\left(u_0^{(l,j)};\mathbf{y},\beta\right)$ and $\rho\left(\mathbf{A}_{0,l}-\mathbf{A}_0\right)+C_{F\Omega}\left\|B_{0,l}-B_0\right\|_{L^\infty(\Omega)}$, which both tend to zero as l and j increase.*

To conclude, we insert the approximation error of the homogenized coefficients from Propositions 4.6 and 4.15 and the combined modelling/discretization error of the homogenized problem from Proposition 4.16 into the estimate of Theorem 4.14 and get the following total error estimate:

Theorem 4.18 (Generalized total error majorant). *Let $\mathbf{A}_\varepsilon$, $\mathbf{b}_\varepsilon$ and c_ε be defined by (3.3) and conditions (3.1) and (3.2) be satisfied. Further, let $\mathbf{A}_0$, B_0 and c_0 be defined by (4.3) and $\mathbf{A}_{0,l}$ and $B_{0,l}$ be defined by (4.6) and (4.7), respectively. We assume that $f\in L^2(\Omega)$, u_ε is the exact solution of (4.1), $u_0^{(l,j)}$ is an approximation of (4.8) and $\widehat{N}_k^{(l)}$, for $k=1,\dots,d$, is an approximation of (4.2). Further, we assume that the approximation $\widetilde{w}_{1,\varepsilon}^{(l,j)}$ is defined by (4.10). Then,*

$$\begin{aligned}\left\|\nabla\left(u_\varepsilon-\widetilde{w}_{1,\varepsilon}^{(l,j)}\right)\right\|_{\mathbf{A}_\varepsilon}&\leq\mathcal{M}_{\mathrm{tot}}\left(\widehat{\mathbf{N}}^{(l)},\mathbf{A}_{0,l},B_{0,l},u_0^{(l,j)},\widetilde{w}_{1,\varepsilon}^{(l,j)}\right)\\&\coloneqq\left(C_1+C_2\delta_{l,j}\right)\frac{\sqrt{\widehat{\alpha}^{\mathrm{cont}}}}{|\widehat{\Pi}|^{1/2}}\sqrt{\sum_{k=1}^d\mathcal{M}_{\mathrm{disc}}^2\left(\widehat{N}_k^{(l)};\widehat{\mathbf{y}},\widehat{\beta}\right)}\\&\quad+\left(C_1+C_2\delta_{l,j}\right)C_{F\Omega}\frac{\left\|\mathbf{b}_\varepsilon\right\|_{L^\infty(\Omega)}}{\sqrt{\widehat{\alpha}^{\mathrm{ell}}}\,|\widehat{\Pi}|^{1/2}}\max_{1\leq k\leq d}\left|\mathcal{M}_{\mathrm{disc}}\left(\widehat{N}_k^{(l)};\widehat{\mathbf{y}},\widehat{\beta}\right)\right|\\&\quad+C_2\kappa_l\mathcal{M}_{\mathrm{disc}}\left(u_0^{(l,j)};\mathbf{y},\beta\right)+\frac{1}{\sqrt{\alpha_\varepsilon^{\mathrm{ell}}}}\left\|\mathbf{A}_{0,l}\nabla u_0^{(l,j)}-\mathbf{A}_\varepsilon\nabla\widetilde{w}_{1,\varepsilon}^{(l,j)}\right\|_{L^2(\Omega)}\\&\quad+\frac{C_{F\Omega}}{\sqrt{\alpha_\varepsilon^{\mathrm{ell}}}}\left(\left\|\left\langle B_{0,l},\nabla u_0^{(l,j)}\right\rangle-\left\langle\mathbf{b}_\varepsilon,\nabla\widetilde{w}_{1,\varepsilon}^{(l,j)}\right\rangle\right\|_{L^2(\Omega)}+\left\|c_0u_0^{(l,j)}-c_\varepsilon\widetilde{w}_{1,\varepsilon}^{(l,j)}\right\|_{L^2(\Omega)}\right),\end{aligned}$$

with

$$\begin{aligned}C_1&\coloneqq\frac{C_{F\Omega}}{\sqrt{\alpha_\varepsilon^{\mathrm{ell}}}\sqrt{\alpha_0^{\mathrm{ell}}}}\left\|f\right\|_{L^2(\Omega)},\qquad C_2\coloneqq\frac{1}{\sqrt{\alpha_\varepsilon^{\mathrm{ell}}}\sqrt{\alpha_0^{\mathrm{ell}}}}\left(\alpha_{0,l}^{\mathrm{cont}}+C_{F\Omega}\left\|B_{0,l}\right\|_{L^\infty(\Omega)}+C_{F\Omega}^2\left\|c_0\right\|_{L^\infty(\Omega)}\right),\\\delta_{l,j}&\coloneqq\frac{1}{\sqrt{\alpha_0^{\mathrm{ell}}}\sqrt{\alpha_{0,l}^{\mathrm{ell}}}}\left(\mathcal{M}_{\mathrm{disc}}\left(u_0^{(l,j)};\mathbf{y},\beta\right)+\left\|\nabla u_0^{(l,j)}\right\|_{\mathbf{A}_{0,l}}\right).\end{aligned}$$

Hence, the total error majorant is a combination of the discretization majorant $\mathcal{M}_{\mathrm{disc}}\left(\widehat{N}_k^{(l)}\right)$ of the cell problems from Proposition 4.5 and the discretization majorant of the homogenized problem $\mathcal{M}_{\mathrm{disc}}\left(u_0^{(l,j)}\right)$ from Proposition 4.16 and three additional computable terms for each coefficient, measuring to some extent the error of the two scale approximation.

Remark 4.19 (Computability). *The total error majorant from Theorem 4.18 is fully computable, due to Remark 4.12 and the following arguments. Since the vector $\mathbf{b}_\varepsilon$ is given, we can explicitly compute $\left\|\mathbf{b}_\varepsilon\right\|_{L^\infty(\Omega)}$. Further, for the constant vector $B_{0,l}$ and the constant coefficient c_0, we can directly compute $\left\|B_{0,l}\right\|_{L^\infty(\Omega)}$ and $\left\|c_0\right\|_{L^\infty(\Omega)}$.*

5 Implementation

In this chapter we will give a brief overview about the implementation which was done in Matlab. For the triangulation we implemented a so called `Mesh` class, which is able to refine and derefine an initial mesh in two dimensions, where the important details are explained in Section 5.1. For the finite element method we implemented a `FEM` class, where the important details and routines are explained in Section 5.2.

5.1 Mesh Refinement and Derefinement

For the adaptive mesh refinement and coarsening we used and revised the algorithm of [30]. The method used is **newest vertex bisection** as described for instance in [7] or [11]. The described method has the advantage that we do not need a tree structure of the triangles.

We consider an admissible mesh $\mathcal{T}$ of Ω, which consists of triangles $T \in \mathcal{T}$, nodes $\mathbf{x} \in \mathcal{N}_\mathcal{T}$ and edges $\varepsilon \in \mathcal{E}_\mathcal{T}$. The nodes and edges are numbered such that $\mathbf{x}_T^i$ is opposite to ε_T^i for $i = 1, 2, 3$. The orientation of the nodes can be counter-clockwise or clockwise, which affects the sign of the area $|T|$. An edge has the nodes $\mathbf{x}_\varepsilon^i$, $i = 1, 2$. Further, an edge is either an inner edge and belongs to the triangles T_ε^i, $i = 1, 2$, or is a boundary edge and belongs to the triangle T_ε^1, or is an edge on the periodic boundary and belongs to the triangles T_ε^1 and $T_{\varepsilon_{\text{per}}}^1$, where ε_{per} is the edge corresponding to ε due to periodicity.

According to the implementation of [30], each triangle T has one **refinement edge** and the numbering is chosen such that ε_T^1 is the refinement edge.

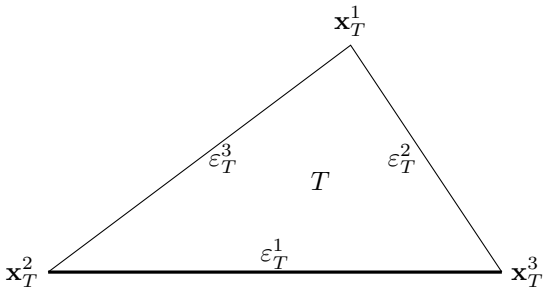


Figure 5.1: A triangle T with nodes $\mathbf{x}_T^i$, $i = 1, 2, 3$, refinement edge ε_T^1 and edges ε_T^2, ε_T^3.

In [7] it is explained, that the refinement edge has to be chosen carefully. In the case of two neighbouring triangles, their common edge should be the refinement edge of both or of neither. If the triangle is a boundary element in the periodic case, the common edge with the corresponding triangle should be the refinement edge of both or of neither due to periodicity.

Additionally, for every triangle we save its **generation** $g_T \in \mathbb{N}_0$. The generation gives the number of ancestors, i.e., if $g_T = 0$ then T is an initial triangle. For every node we save its **flag** $f_\mathbf{x} \in \{-1, 0, 1, 2\}$, giving the information whether it is an inner node $f_\mathbf{x} = 0$ or a node on the Dirichlet boundary $f_\mathbf{x} = 1$, on the Neumann boundary $f_\mathbf{x} = -1$ or on the periodic boundary $f_\mathbf{x} = 2$. In the case of a mixed Dirichlet-Neumann boundary, the Dirichlet node has the priority. Further, for every edge we save its **flag** $f_\varepsilon \in \{-1, 0, 1, 2\}$, giving the information whether it is an inner edge $f_\varepsilon = 0$ or an edge on the Dirichlet boundary $f_\varepsilon = 1$, on the Neumann boundary $f_\varepsilon = -1$ or on the periodic boundary $f_\varepsilon = 2$.

The code considers for each triangle local and global numbering, therefore refinement and derefinement do not need additional data structures.

5.1.1 Marking

The marking scheme is taken from [30], which has the advantage of checking regularity first and then refine or derefine all triangles at once. We only added the periodic case, which results in some small adaptations. In this scheme, the edges of the mesh get a marking, where -1 stands for derefine, 1 for refine and 0 for no derefine or refine.
The marking should lead to an admissible mesh, this means that no hanging nodes and no marking contradictions (e.g., edges of the same triangle marked for refinement and derefinement) should arise. In the periodic case we additionally have to consider that the marking of ε should be the same as ε_{per}.
Given an arbitrary marking m we can always construct a valid marking m_v. This can be done by spreading marks to other edges, removing contradictions and by giving priority to refinement markings.
The following refinement cases can occur, due to different marking:

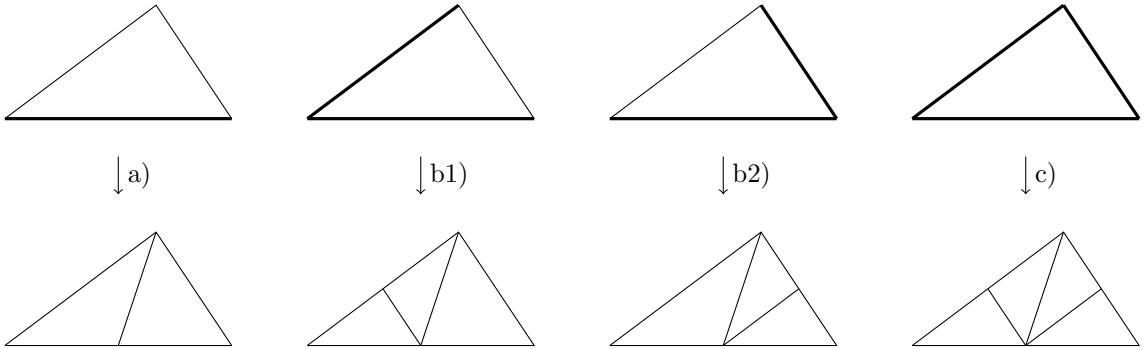


Figure 5.2: Refinement for different markings.

In the case of uniform refinement, we usually mark all edges and hence get a refinement of a triangle as in Figure 5.2 the case c).

Example 5.1. We give here an example of a very coarse mesh, actually an initial mesh with mixed Dirichlet and Neumann boundary condition, in order to illustrate the data structures, see Table 5.1. The global numbering of the thirteen nodes is shown in Figure 5.3.

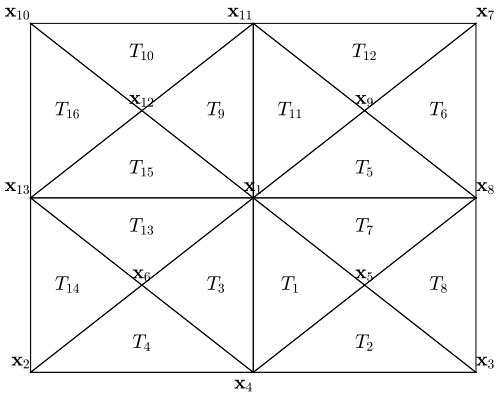


Figure 5.3: Example of a mesh.

The considered mesh consists of thirteen nodes, i.e. $N = 13$, and of sixteen triangles. The `Mesh` class stores a list of nodes, containing the x- and y-coordinate of each `node` according to the global numbering. Further, the `flag` of every node is stored in a vector. Then, there is a list of triangles, which stores in each row the nodes of a `triangle` as indices into `node`. This corresponds to the

local numbering of the geometric nodes. The ordering is also reflected in the sign of the `area`, e.g. the triangle T_1 is oriented counter-clockwise and T_2 clockwise. There is a list of edge indices, which stores in each row the nodes of an `edge` as indices into `node`, e.g. $\varepsilon_9 = [2 \quad 4]$, and a vector giving the flag of each edge, e.g. $f_{\varepsilon_9} = -1$. Then, there is a vector storing each `diameter` h_T.
The Dirichlet and Neumann nodes can easily be calculated from the flags $f_{\mathbf{x}}$ and we know the Dirichlet and Neumann boundary from the flags f_ε. The condition that the Dirichlet node has the priority can be understood like this: The edge ε_9 lies on the Neumann boundary ($f_{\varepsilon_9} = -1$) and the edge $\varepsilon_{11} = [2 \quad 13]$ lies on the Dirichlet boundary ($f_{\varepsilon_{11}} = 1$), since the Dirichlet property has priority we have $f_{\mathbf{x}_2} = 1$.
There are further lists and numbers stored, which are used either for the mesh generation or refinement or are used often and therefore the performance of the code improves.

	mesh.node		flag
i	$x_{i,1}$	$x_{i,2}$	$f_{\mathbf{x}}$
1	0.5	0.5	0
2	0	0	1
3	1	0	1
4	0.5	0	-1
5	0.75	0.25	0
6	0.25	0.25	0
7	1	1	1
8	1	0.5	1
9	0.75	0.75	0
10	0	1	1
11	0.5	1	-1
12	0.25	0.75	0
13	0	0.5	1

	mesh.triangle			mesh.area
i	$\mathbf{x}^1_{T_i}$	$\mathbf{x}^2_{T_i}$	$\mathbf{x}^3_{T_i}$	$\lvert T_i\rvert$
1	5	1	4	1/16
2	5	3	4	-1/16
3	6	1	4	-1/16
4	6	2	4	1/16
5	9	1	8	1/16
6	9	7	8	-1/16
7	5	1	8	-1/16
8	5	3	8	1/16
9	12	1	11	1/16
10	12	10	11	-1/16
11	9	1	11	-1/16
12	9	7	11	1/16
13	6	1	13	1/16
14	6	2	13	-1/16
15	12	1	13	-1/16
16	12	10	13	1/16

	dirichlet	
i	$\mathbf{x}_{i,1}$	$\mathbf{x}_{i,2}$
1	0	0
2	1	0
3	1	1
4	1	0.5
5	0	1
6	0	0.5

	neumann	
i	$\mathbf{x}_{i,1}$	$\mathbf{x}_{i,2}$
1	0.5	0
2	0.5	1

Table 5.1: Data structure.

5.2 Finite Element Method

The computation of a numerical approximation of the elliptic partial differential equation is done in several steps: First, we construct a mesh that partitions Ω, then, we assemble the system matrix $\mathbf{L}$ and the right-hand side vector $\mathbf{f}$. Further, we solve the linear system and finally compute the a priori or a posteriori error estimate.

5.2.1 Assembling of the System Matrix and the Right-Hand Side

In Section 2.3 we derived the system of equations (2.30): $\mathbf{Lu} = \mathbf{f}$, where

$$l_{i,j} = a(\psi_j, \psi_i), \quad f_i = l(\psi_i), \quad \forall i, j = 1, \dots, N.$$

First, we want to compute the system matrix $\mathbf{L}$. We only explain the procedure for the part with diffusion matrix $\mathbf{A}$ and remark how to adapt for the lower order terms $\mathbf{b}$ and c later, hence we consider the following integral:

$$a(\psi_j, \psi_i) = \int_\Omega \langle \mathbf{A}\nabla\psi_j, \nabla\psi_i\rangle \, \mathrm{d}\mathbf{x} = \sum_{T\in\mathcal{T}_h} \int_T \langle \mathbf{A}\nabla\psi_j, \nabla\psi_i\rangle \, \mathrm{d}\mathbf{x} =: \sum_{T\in\mathcal{T}_h} a_T(\psi_j, \psi_i). \tag{5.1}$$

The basis functions ψ_i are defined over the s nodal basis functions $\widehat{\psi}_i$ defined on T_{ref}. It is clear, that they have a small support, which is exactly ω_i as defined in (2.37). The integral then only has

to be computed when T lies in the supports of ψ_i and ψ_j, since it vanishes otherwise. It is most efficient to assemble the system matrix elementwise, i.e., for every $T \in \mathcal{T}_h$ we derive a $s \times s$ matrix $(a_T(\psi_j, \psi_i))_{i,j}$, which contains all the combinations of ψ_i and ψ_j for $i, j = 1, 2, \ldots, s$. Then, those submatrices are added up for every node $\mathbf{x}_i$, $i = 1, \ldots, N$, corresponding to the triangle that contain this node. In this context we speak of a local degree of freedom s and a global degree of freedom N. In Example 5.1 the support of $\mathbf{x}_{13}$ (for Dirichlet or Neumann boundary conditions) is the union of $\{T_{13}, T_{14}, T_{15}, T_{16}\}$. We assemble the same system matrix for Dirichlet and for Neumann boundary conditions, but the system will be solved only for the global degree of freedoms.
The basis functions, in this context we consider linear ones, are defined on T_{ref} with the nodes $(0,0)$, $(1,0)$, $(1,1)$ as:

$$\widehat{\psi}_1(\hat{\mathbf{x}}) = 1 - \hat{x}_1, \quad \widehat{\psi}_2(\hat{\mathbf{x}}) = \hat{x}_1 - \hat{x}_2, \quad \widehat{\psi}_3(\hat{\mathbf{x}}) = \hat{x}_2, \tag{5.2}$$

with the gradient

$$\nabla_{\hat{\mathbf{x}}}\widehat{\psi}_1(\hat{\mathbf{x}}) = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \quad \nabla_{\hat{\mathbf{x}}}\widehat{\psi}_2(\hat{\mathbf{x}}) = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \nabla_{\hat{\mathbf{x}}}\widehat{\psi}_3(\hat{\mathbf{x}}) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{5.3}$$

Therefore, we use the affine mapping χ_T defined in (2.33) to transform the integral onto T_{ref}:

$$\begin{aligned} a_T(\psi_j, \psi_i) &= |\det(\mathrm{D}\chi_T)| \int_{T_{\text{ref}}} \left\langle \mathbf{A}(\chi_T(\hat{\mathbf{x}}))(\mathrm{D}\chi_T)^{-\top} \nabla_{\hat{\mathbf{x}}}\psi_j(\chi_T(\hat{\mathbf{x}})), (\mathrm{D}\chi_T)^{-\top} \nabla_{\hat{\mathbf{x}}}\psi_i(\chi_T(\hat{\mathbf{x}})) \right\rangle \mathrm{d}\hat{\mathbf{x}} \\ &= |2|T|| \int_{T_{\text{ref}}} \left\langle \mathbf{A}(\chi_T(\hat{\mathbf{x}}))(\mathrm{D}\chi_T)^{-\top} \nabla_{\hat{\mathbf{x}}}\widehat{\psi}_j(\hat{\mathbf{x}}), (\mathrm{D}\chi_T)^{-\top} \nabla_{\hat{\mathbf{x}}}\widehat{\psi}_i(\hat{\mathbf{x}}) \right\rangle \mathrm{d}\hat{\mathbf{x}}, \end{aligned}$$

where the Jacobian is defined as in (2.34), with the convention $(\mathrm{D}\chi_T)^{-\top} := \left((\mathrm{D}\chi_T)^{\top}\right)^{-1}$, and we used the equation (2.35) for the determinant of the Jacobian. The absolute value of the area $|T|$ is important, since the signed area of T can be negative according to the orientation of the nodes. Furthermore, the relation

$$\nabla_{\hat{\mathbf{x}}}\widehat{\psi} = (\mathrm{D}\chi_T)^{\top} \nabla\psi$$

was used.
In addition, we can use simplex coordinates to transform the integral over T_{ref} to an integral over the unit square $[0,1]^2$, see e.g. [23]. This is achieved with an affine function F defined by:

$$\begin{aligned} F : [0,1]^2 &\to T_{\text{ref}} \\ (\xi, \eta) &\to (\hat{x}_1, \hat{x}_2) = F(\xi, \eta) := (\xi, \xi\eta). \end{aligned}$$

The determinant of the Jacobian matrix of F is

$$\det(\mathrm{D}F) = \begin{vmatrix} 1 & 0 \\ \eta & \xi \end{vmatrix} = \xi.$$

Again we have a relation

$$\nabla_{(\xi,\eta)}\widehat{\psi} = (\mathrm{D}F)^{\top} \nabla_{\hat{\mathbf{x}}}\widehat{\psi}.$$

The integral now takes the form

$$\begin{aligned} a_T(\psi_j, \psi_i) = |2|T|| \int_0^1 \int_0^1 \xi \Big\langle \mathbf{A}(\chi_T(\xi, \xi\eta))(\mathrm{D}\chi_T)^{-\top} (\mathrm{D}F)^{-\top} \nabla_{(\xi,\eta)}\widehat{\psi}_j(\xi, \xi\eta), \\ (\mathrm{D}\chi_T)^{-\top} (\mathrm{D}F)^{-\top} \nabla_{(\xi,\eta)}\widehat{\psi}_i(\xi, \xi\eta) \Big\rangle \mathrm{d}\xi \, \mathrm{d}\eta. \end{aligned}$$

The routine should be very flexible with respect to the basis functions $\widehat{\psi}_i$ and should allow matrices (or coefficients) depending on $\mathbf{x}$. Therefore, we have to approximate the integral over the unit square by a quadrature formula. We use the tensor product of a standard one dimensional Gauss quadrature formula with abscissas $(x_{i,n_Q})_{i=1}^{n_Q}$ and weights $(w_{i,n_Q})_{i=1}^{n_Q}$, which is exact for polynomials of order at most $2n_Q - 1$. The computation of the weights and abscissas is explained, e.g. in [18]. As an example, see Figure 5.4, where the abscissas of the unit square are transformed onto a simple mesh $\mathcal{T}_h$.

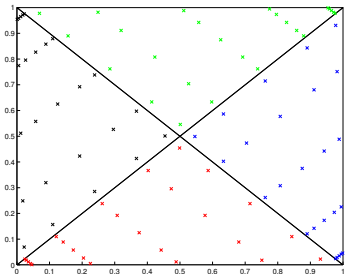


Figure 5.4: Quadrature points for a mesh with 4 elements.

Finally, what we compute is:

$$\begin{aligned} a_T(\psi_j,\psi_i) = |2|T|| \sum_{k=1}^{n_Q}\sum_{l=1}^{n_Q} w_{k,n_Q} w_{l,n_Q} x_{k,n_Q} \\ \Big\langle \mathbf{A}\left(\chi_T(x_{k,n_Q}, x_{k,n_Q}x_{l,n_Q})\right) (\mathrm{D}\chi_T)^{-\top} (\mathrm{D}F)^{-\top}_{k,l} \nabla_{(x_{k,n_Q},x_{l,n_Q})} \widehat{\psi}_j(x_{k,n_Q}, x_{k,n_Q}x_{l,n_Q}), \\ (\mathrm{D}\chi_T)^{-\top} (\mathrm{D}F)^{-\top}_{k,l} \nabla_{(x_{k,n_Q},x_{l,n_Q})} \widehat{\psi}_i(x_{k,n_Q}, x_{k,n_Q}x_{l,n_Q})\Big\rangle, \end{aligned}$$

with

$$(\mathrm{D}\chi_T)^{-\top} = \frac{1}{2|T|}\begin{bmatrix} y_C - y_B & y_A - y_B \\ x_B - x_C & x_B - x_A \end{bmatrix} \quad \text{and} \quad (\mathrm{D}F)^{-\top}_{k,l} = \frac{1}{x_{k,n_Q}}\begin{bmatrix} x_{k,n_Q} & -x_{l,n_Q} \\ 0 & 1 \end{bmatrix}.$$

With additional coefficients $\mathbf{b}$ and c, we have

$$\begin{aligned} a(\psi_j,\psi_i) &= \int_\Omega \left\{\langle \mathbf{A}\nabla\psi_j, \nabla\psi_i\rangle + \langle \mathbf{b}, \nabla\psi_j\rangle \psi_i + c\psi_j\psi_i\right\} \mathrm{d}\mathbf{x} \\ &= \underbrace{\sum_{T\in\mathcal{T}_h} \int_T \langle \mathbf{A}\nabla\psi_j, \nabla\psi_i\rangle \,\mathrm{d}\mathbf{x}}_{\text{part 1}} + \underbrace{\sum_{T\in\mathcal{T}_h} \int_T \langle \mathbf{b}, \nabla\psi_j\rangle \psi_i \,\mathrm{d}\mathbf{x}}_{\text{part 2}} + \underbrace{\sum_{T\in\mathcal{T}_h} \int_T c\psi_j\psi_i \,\mathrm{d}\mathbf{x}}_{\text{part 3}}. \end{aligned} \tag{5.4}$$

Part 2 can be computed exactly as part 1, with the only difference that the Jacobians and their determinant appear only once instead of twice. This term is unsymmetric, thus, one has to be careful with the indices i and j during implementation. In particular, the multiplicative constant is in this case $\frac{|2|T||}{2|T|} = \pm 1$, so it changes the sign according to the orientation of the triangle. Part 3 is even more simpler, since no Jacobians and determinants thereof appear.
In the case of a Neumann boundary condition for $c = 0$ and in the case of periodic boundary condition, we use Lagrange multipliers to get a solution up to a constant. The additional term for the bilinear form then is:

$$\int_\Omega u \,\mathrm{d}\mathbf{x} \int_\Omega \psi_i \,\mathrm{d}\mathbf{x} = \sum_{j=1}^{N} u_j \int_\Omega \psi_j \,\mathrm{d}\mathbf{x} \int_\Omega \psi_i \,\mathrm{d}\mathbf{x},$$

which can be computed as explained before. For periodic boundary conditions we have to consider the special support of the basis functions, see Section 5.2.3.
Similarly, we compute the right-hand side vector:

$$\begin{aligned} l(\psi_i) &= \sum_{T\in\mathcal{T}_h} \int_T f\psi_i \,\mathrm{d}\mathbf{x} \\ &= \sum_{T\in\mathcal{T}_h} |2|T|| \sum_{k=1}^{n_Q}\sum_{l=1}^{n_Q} w_{k,n_Q} w_{l,n_Q} x_{k,n_Q} f\left(\chi_T(x_{k,n_Q}, x_{k,n_Q}x_{l,n_Q})\right) \widehat{\psi}_i(x_{k,n_Q}, x_{k,n_Q}x_{l,n_Q}). \end{aligned} \tag{5.5}$$

In the case of a Neumann boundary condition we have the linear form

$$l(\psi_i) = \int_\Omega f\psi_i \,\mathrm{d}\mathbf{x} + \int_\Gamma g\psi_i \,\mathrm{d}s,$$

thus we have an additional term, which corresponds to a one-dimensional integration where, due to the orientation, again a constant ± 1 appears.

5.2.2 Solving the Linear System

After assembling $\mathbf{L}$ and $\mathbf{f}$ we want to solve the linear system $\mathbf{Lu} = \mathbf{f}$ to get $\mathbf{u}$. In the case of the Dirichlet boundary condition $u = g$ on Γ, the value of $\mathbf{u}$ on the boundary is given by the function g. Therefore, we do not solve the entire linear system. Assume that we have n inner nodes and $N - n$ boundary nodes, then the global degree of freedom is n and we solve

$$\sum_{j=1}^{n} u_j a(\psi_j, \psi_i) = l(\psi_i) - \sum_{j=n+1}^{N} u_j a(\psi_j, \psi_i) \quad \text{for } i = 1, 2, \ldots, n, \tag{5.6}$$

where $u_j = g(\mathbf{x}_j)$ for the boundary nodes $\mathbf{x}_j$, $j = n+1, \ldots, N$. For Neumann boundary conditions, we solve the entire system and the global degree of freedom is N.
The system matrix $\mathbf{L}$ is sparse, due to the small support of the basis functions. Hence, the work to set up the system matrix is linear with respect to the number of unknowns N.

Remark 5.1 (Fast implementation). *The implementation (mesh refinement, assembling and solving the system, etc.) was done in Matlab. In order to solve interesting examples numerically, one has to consider certain speed up techniques. One should try to avoid for-loops by vectorizing every step, important Matlab functions for that are* `reshape`, `repmat`, `meshgrid`, `sparse` *and* `accumarray`. *The possible speed up via fast implementation is explained in detail, e.g. in [17].*

5.2.3 Periodic Boundary Condition

Periodic boundary conditions require a specific treatment which is explained next. Due to the periodicity we only have some degree of freedoms on the boundary and therefore, our linear system should carefully be reduced. Consider the following example (Figure 5.5): All points where we solve the system, i.e. the degree of freedoms, are plotted as a white circle. The matrix entry of a grid point on the left boundary of the domain, according to the periodicity, should be the same as the one on the right boundary.

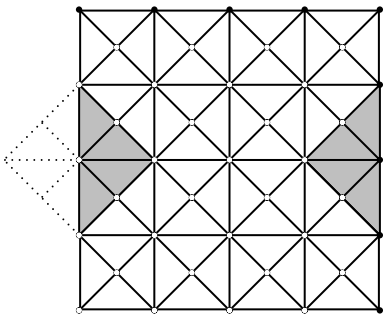


Figure 5.5: Support of a periodic basis function (in gray).

First of all, we have to modify the adaptive refinement strategy such that the points on the left boundary correspond to the right boundary, meaning that if we get an additional point due to refinement on the left, we also have to add it on the right and other necessary points.
Secondly, we add up element matrices, therefore we have to modify which element matrices we take. Since the point on the right is no degree of freedom, it is somehow natural to add its support to the support for the white circle on the left, see Figure 5.5. If you would wrap this square mesh to a cylinder, it gets clear that this is the way to create periodic basis functions. Consider again Example 5.1, for a Dirichlet or Neumann boundary condition the support is given by `mesh.triangle` (for every node we find immediately which triangles contain this node) and this would be only half of the support in the periodic case, as depicted in the figure. Therefore we additionally store a list

called `support`, which replaces the nodes that are no degree of freedom with their corresponding degree of freedom due to periodicity. In Example 5.1 the support of $\mathbf{x}_{13}$ would then be the union of $\{T_5, T_6, T_7, T_8, T_{13}, T_{14}, T_{15}, T_{16}\}$ and the index 8 would be replaced by the index 13.
We solve a reduced linear system, meaning only in the degrees of freedom inside the domain and on the left and the lower boundary. Therefore, we store a list of `freenodes` that contains all the degree of freedoms, i.e. all the white circles. The values of the finite element solution on the other boundary nodes obtain the values of their corresponding nodes on the left or lower boundary.

5.2.4 Finite Elements

In Chapter 2 we defined the finite element space and the Lagrange basis. We already mentioned, that the linear basis functions on T_{ref} with the nodes $(0,0)$, $(1,0)$, $(1,1)$ are:

$$\widehat{\psi}_1(\hat{\mathbf{x}}) = 1 - \hat{x}_1, \quad \widehat{\psi}_2(\hat{\mathbf{x}}) = \hat{x}_1 - \hat{x}_2, \quad \widehat{\psi}_3(\hat{\mathbf{x}}) = \hat{x}_2.$$

The local degree of freedom is in this case $s = 3$ and the geometric nodes (the nodes that define the triangle) are equal to the algebraic nodes. The algebraic nodes are the s nodes defined in Remark 2.23 with $s = \frac{(t+1)(t+2)}{2}$, where t is the order.
For order two, we get six degree of freedoms, i.e. six algebraic nodes. The quadratic basis functions can also easily be derived as:

$$\widehat{\psi}_1(\hat{\mathbf{x}}) = 1 - 3\hat{x}_1 + 2\hat{x}_1^2, \quad \widehat{\psi}_2(\hat{\mathbf{x}}) = -\hat{x}_1 + \hat{x}_2 - 4\hat{x}_1\hat{x}_2 + 2\hat{x}_1^2 + 2\hat{x}_2^2, \quad \widehat{\psi}_3(\hat{\mathbf{x}}) = -\hat{x}_2 + 2\hat{x}_2^2,$$
$$\widehat{\psi}_4(\hat{\mathbf{x}}) = 4\hat{x}_1\hat{x}_2 - 4\hat{x}_2^2, \quad \widehat{\psi}_5(\hat{\mathbf{x}}) = 4\hat{x}_2 - 4\hat{x}_1\hat{x}_2, \quad \widehat{\psi}_6(\hat{\mathbf{x}}) = 4\hat{x}_1 - 4\hat{x}_2 + 4\hat{x}_1\hat{x}_2 - 4\hat{x}_1^2.$$

Considering the data structure and Example 5.1, this means that we define a FEM class with additional information. We store a new list `fem.node` with all algebraic nodes, which contains 41 nodes in this particular case. Further, we store a list of triangles, which is the original one extended with the indices of the additional nodes, e.g. $T_1 = [5, 1, 4, 14, 15, 16]$.
For the implementation, we need to store the basis functions and their gradients on the reference triangle. We want to have a general procedure, where we can decide a polynomial degree t and then, the basis functions are derived once and stored appropriately, this can be done with the following considerations.
For general degree t and s degree of freedoms, we first have to define a numbering strategy and store the algebraic nodes of the reference triangle according to it: The first three nodes are chosen arbitrarily but counter-clockwise, then we continue numbering on the edge opposite of the first node. For $t \geq 3$, we will also have degree of freedoms inside the triangle, there the numbering continues again as before, see Figure 5.6 for an illustration.

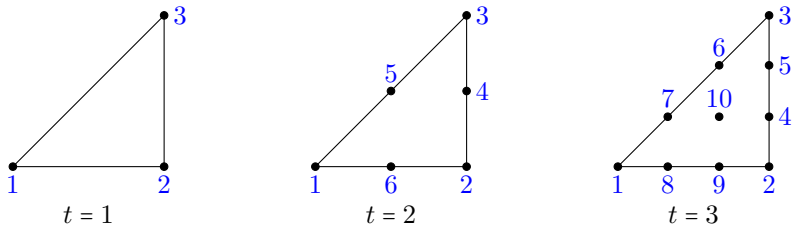


Figure 5.6: Numbering of the local degree of freedoms.

The basis functions can be computed by an explicit formula with the idea from [15]. For that, we first denote the s nodes of degree t on the reference triangle by:

$$\Theta_t := \left\{ \frac{\boldsymbol{\alpha}}{t} \,\middle|\, \boldsymbol{\alpha} = (\alpha_1, \alpha_2) \in \mathbb{N}_0^2 \text{ with } \alpha_2 - \alpha_1 \leq 0 \text{ and } \alpha_1, \alpha_2 \leq t \right\},$$

with the ordering as explained before. Then, the Lagrange basis function for the node $\hat{\mathbf{z}}_l = (\hat{z}_1, \hat{z}_2) = \frac{1}{t}(\alpha_1, \alpha_2) \in \Theta_t$ is given by

$$\widehat{\psi}_{t,l}(\hat{x}_1, \hat{x}_2) = \prod_{j=0}^{\alpha_1-\alpha_2-1} \frac{j - t(\hat{x}_1 - \hat{x}_2)}{j - t(\hat{z}_1 - \hat{z}_2)} \prod_{j=0}^{\alpha_2-1} \frac{j - t\hat{x}_2}{j - t\hat{z}_2} \prod_{j=\alpha_1+1}^{t} \frac{j - t\hat{x}_1}{j - t\hat{z}_1},$$

for $l = 1, 2, \ldots, s$.
Those basis functions fulfil the Lagrange condition:

$$\widehat{\psi}_{t,l}(\hat{\mathbf{z}}_j) = \delta_{l,j} \quad \text{for } l, j = 1, 2, \ldots, s,$$

which means that the functions are equal to one for exactly one node and zero at all the other nodes. Finally, we use this formula for defining the shape functions $\widehat{\psi}_{t,l}$, $l = 1, 2, \ldots, s$, and compute their gradient.
In the quadratic case, we have the combinations

$$\boldsymbol{\alpha} = \{(0,0), (2,0), (2,2), (2,1), (1,1), (1,0)\}$$

and we get, e.g.,

$$\widehat{\psi}_{2,4}(\hat{x}_1, \hat{x}_2) = \frac{0 - 2(\hat{x}_1 - \hat{x}_2)}{0 - 2\left(1 - \frac{1}{2}\right)} \frac{0 - 2\hat{x}_2}{0 - 2\frac{1}{2}} = 4\hat{x}_1\hat{x}_2 - 4\hat{x}_2^2.$$

5.2.5 Majorant

The concept of a posteriori error estimation was explained in Section 2.4, we now consider the general implementation of the developed majorants.
The first idea for the choice of the flux variable $\mathbf{y}$ in the error majorant of Theorem 2.37 is $\mathbf{y} = \mathbf{A}\nabla v_h$, where v_h is our finite element approximation. Then, as we mentioned in Section 2.5, the flux $\mathbf{y}$ would not be in $H(\Omega, \mathrm{div})$. Therefore, some extra effort is required to get a good reconstruction of the flux. One strategy to overcome this problem is to post-process the flux with a gradient recovery procedure, as explained in Section 2.5. Then, one further has to employ some minimization steps (c.f. [29]), to get an efficient estimate. In the following, we explain the minimization procedure to improve a first approximation of the flux (see [26] or [24]).
By squaring the first and the second term of the majorant defined in Theorem 2.37 and using Young's inequality we arrive at

$$\begin{aligned}\mathcal{M}^2(v, \mathbf{y}; \beta, \gamma) \coloneqq{}& (1+\beta)\|\mathbf{y} - \mathbf{A}\nabla v\|^2_{\mathbf{A}^{-1}} \\ &+ \frac{1+\beta}{\beta}\frac{1}{\alpha^{\mathrm{ell}}}\left((1+\gamma)C^2_{F\Omega}\|r_\Omega(v,\mathbf{y})\|^2_{L^2(\Omega)} + \frac{1+\gamma}{\gamma}C^2_{T\Gamma_N}\|g_N - \langle \mathbf{y}, \mathbf{n}\rangle\|^2_{L^2(\Gamma_N)}\right). \quad (5.7)\end{aligned}$$

We start by minimizing the right-hand side of (5.7) with respect to γ:

$$\frac{\mathrm{d}\mathcal{M}^2(v, \mathbf{y}; \beta, \gamma)}{\mathrm{d}\gamma} = \frac{1+\beta}{\beta}\frac{C^2_{F\Omega}}{\alpha^{\mathrm{ell}}}\|r_\Omega(v,\mathbf{y})\|^2_{L^2(\Omega)} + \frac{1+\beta}{\beta}\frac{-1}{\gamma^2}\frac{C^2_{T\Gamma_N}}{\alpha^{\mathrm{ell}}}\|g_N - \langle \mathbf{y}, \mathbf{n}\rangle\|^2_{L^2(\Gamma_N)} = 0,$$

which leads to the minimizer

$$\hat{\gamma} = \sqrt{\frac{C^2_{T\Gamma_2}\|g_N - \langle \mathbf{y}, \mathbf{n}\rangle\|^2_{L^2(\Gamma_N)}}{C^2_{F\Omega}\|r_\Omega(v,\mathbf{y})\|^2_{L^2(\Omega)}}} = \frac{C_{T\Gamma_N}\|g_N - \langle \mathbf{y}, \mathbf{n}\rangle\|_{L^2(\Gamma_N)}}{C_{F\Omega}\|r_\Omega(v,\mathbf{y})\|_{L^2(\Omega)}}. \quad (5.8)$$

Consequently, this choice of $\hat{\gamma}$ gives the majorant

$$\begin{aligned}\mathcal{M}^2_\gamma(v, \mathbf{y}; \beta) \coloneqq{}& (1+\beta)\|\mathbf{y} - \mathbf{A}\nabla v\|^2_{\mathbf{A}^{-1}} \\ &+ \frac{1+\beta}{\beta}\frac{1}{\alpha^{\mathrm{ell}}}\left(C_{F\Omega}\|r_\Omega(v,\mathbf{y})\|_{L^2(\Omega)} + C_{T\Gamma_N}\|g_N - \langle \mathbf{y}, \mathbf{n}\rangle\|_{L^2(\Gamma_N)}\right)^2. \quad (5.9)\end{aligned}$$

Now, we fix γ and minimize the majorant (5.9) with respect to β:

$$\frac{\mathrm{d}\mathcal{M}_\gamma^2(v,\mathbf{y};\beta)}{\mathrm{d}\beta} = \|\mathbf{y}-\mathbf{A}\nabla v\|_{\mathbf{A}^{-1}}^2 - \frac{1}{\beta^2}\frac{1}{\alpha^{\mathrm{ell}}}\left(C_{F\Omega}\|r_\Omega(v,\mathbf{y})\|_{L^2(\Omega)} + C_{T\Gamma_N}\|g_N - \langle \mathbf{y},\mathbf{n}\rangle\|_{L^2(\Gamma_N)}\right)^2 = 0,$$

which leads to the optimal

$$\begin{aligned}\hat{\beta} &= \sqrt{\frac{\frac{1}{\alpha^{\mathrm{ell}}}\left(C_{F\Omega}\|r_\Omega(v,\mathbf{y})\|_{L^2(\Omega)} + C_{T\Gamma_N}\|g_N - \langle \mathbf{y},\mathbf{n}\rangle\|_{L^2(\Gamma_N)}\right)^2}{\|\mathbf{y}-\mathbf{A}\nabla v\|_{\mathbf{A}^{-1}}^2}} \\ &= \frac{C_{F\Omega}\|r_\Omega(v,\mathbf{y})\|_{L^2(\Omega)} + C_{T\Gamma_N}\|g_N - \langle \mathbf{y},\mathbf{n}\rangle\|_{L^2(\Gamma_N)}}{\sqrt{\alpha^{\mathrm{ell}}}\|\mathbf{y}-\mathbf{A}\nabla v\|_{\mathbf{A}^{-1}}}.\end{aligned} \tag{5.10}$$

Finally, we fix β and minimize with respect to $\mathbf{y}$. For that, we substitute $\mathbf{y}$ by $\mathbf{y}+t\boldsymbol{\mu}$ for $t\in\mathbb{R}$ and $\boldsymbol{\mu}\in H(\Omega,\mathrm{div})$. Now we take the derivative of

$$\begin{aligned}\mathcal{M}_{\beta,\gamma}^2(v,\mathbf{y}+t\boldsymbol{\mu}) = (1+\beta)\|\mathbf{y}+t\boldsymbol{\mu}-\mathbf{A}\nabla v\|_{\mathbf{A}^{-1}}^2 + \frac{1+\beta}{\beta}\frac{C_{T\Gamma_N}^2}{\alpha^{\mathrm{ell}}}\frac{1+\gamma}{\gamma}\|g_N - \langle \mathbf{y},\mathbf{n}\rangle - t\langle \boldsymbol{\mu},\mathbf{n}\rangle\|_{L^2(\Gamma_N)}^2 \\ + \frac{1+\beta}{\beta}\frac{C_{F\Omega}^2}{\alpha^{\mathrm{ell}}}(1+\gamma)\|f - \langle \mathbf{b},\nabla v\rangle - cv + \mathrm{div}\,\mathbf{y} + t\,\mathrm{div}\,\boldsymbol{\mu}\|_{L^2(\Omega)}^2\end{aligned}$$

with respect to t:

$$\begin{aligned}\frac{\mathrm{d}\mathcal{M}_{\beta,\gamma}^2(v,\mathbf{y}+t\boldsymbol{\mu})}{\mathrm{d}t} &= (1+\beta)\int_\Omega 2\left\langle \mathbf{A}^{-1}(\mathbf{y}+t\boldsymbol{\mu}-\mathbf{A}\nabla v),\boldsymbol{\mu}\right\rangle \mathrm{d}\mathbf{x} \\ &\quad + \frac{1+\beta}{\beta}\frac{C_{T\Gamma_N}^2}{\alpha^{\mathrm{ell}}}\frac{1+\gamma}{\gamma}\int_{\Gamma_N} 2\left(g_N - \langle \mathbf{y},\mathbf{n}\rangle - t\langle \boldsymbol{\mu},\mathbf{n}\rangle\right)\langle \boldsymbol{\mu},\mathbf{n}\rangle \,\mathrm{d}s \\ &\quad + \frac{1+\beta}{\beta}\frac{C_{F\Omega}^2}{\alpha^{\mathrm{ell}}}(1+\gamma)\int_\Omega 2\left(f - \langle \mathbf{b},\nabla v\rangle - cv + \mathrm{div}\,\mathbf{y} + t\,\mathrm{div}\,\boldsymbol{\mu}\right)\mathrm{div}\,\boldsymbol{\mu}\,\mathrm{d}\mathbf{x}.\end{aligned}$$

By setting the equation for $t=0$ equal to zero, we get the following system:

$$\begin{aligned}&(1+\beta)\int_\Omega \left\langle \mathbf{A}^{-1}\mathbf{y},\boldsymbol{\mu}\right\rangle \mathrm{d}\mathbf{x} + \frac{1+\beta}{\beta}\frac{C_{F\Omega}^2}{\alpha^{\mathrm{ell}}}(1+\gamma)\int_\Omega \mathrm{div}\,\mathbf{y}\,\mathrm{div}\,\boldsymbol{\mu}\,\mathrm{d}\mathbf{x} \\ &\quad - \frac{1+\beta}{\beta}\frac{C_{T\Gamma_N}^2}{\alpha^{\mathrm{ell}}}\frac{1+\gamma}{\gamma}\int_{\Gamma_N}\langle \mathbf{y},\mathbf{n}\rangle\langle \boldsymbol{\mu},\mathbf{n}\rangle\,\mathrm{d}s \\ &= (1+\beta)\int_\Omega \langle \nabla v,\boldsymbol{\mu}\rangle\,\mathrm{d}\mathbf{x} + \frac{1+\beta}{\beta}\frac{C_{F\Omega}^2}{\alpha^{\mathrm{ell}}}(1+\gamma)\int_\Omega \left(\langle \mathbf{b},\nabla v\rangle + cv - f\right)\mathrm{div}\,\boldsymbol{\mu}\,\mathrm{d}\mathbf{x} \\ &\quad - \frac{1+\beta}{\beta}\frac{C_{T\Gamma_N}^2}{\alpha^{\mathrm{ell}}}\frac{1+\gamma}{\gamma}\int_{\Gamma_N} g_N\langle \boldsymbol{\mu},\mathbf{n}\rangle\,\mathrm{d}s,\end{aligned}$$

which is equivalent to

$$\begin{aligned}&\beta\int_\Omega \left\langle \mathbf{A}^{-1}\mathbf{y},\boldsymbol{\mu}\right\rangle \mathrm{d}\mathbf{x} + \frac{C_{F\Omega}^2}{\alpha^{\mathrm{ell}}}(1+\gamma)\int_\Omega \mathrm{div}\,\mathbf{y}\,\mathrm{div}\,\boldsymbol{\mu}\,\mathrm{d}\mathbf{x} - \frac{C_{T\Gamma_N}^2}{\alpha^{\mathrm{ell}}}\frac{1+\gamma}{\gamma}\int_{\Gamma_N}\langle \mathbf{y},\mathbf{n}\rangle\langle \boldsymbol{\mu},\mathbf{n}\rangle\,\mathrm{d}s \\ &= \beta\int_\Omega \langle \nabla v,\boldsymbol{\mu}\rangle\,\mathrm{d}\mathbf{x} + \frac{C_{F\Omega}^2}{\alpha^{\mathrm{ell}}}(1+\gamma)\int_\Omega \left(\langle \mathbf{b},\nabla v\rangle + cv - f\right)\mathrm{div}\,\boldsymbol{\mu}\,\mathrm{d}\mathbf{x} - \frac{C_{T\Gamma_N}^2}{\alpha^{\mathrm{ell}}}\frac{1+\gamma}{\gamma}\int_{\Gamma_N} g_N\langle \boldsymbol{\mu},\mathbf{n}\rangle\,\mathrm{d}s,\end{aligned}$$

$\forall \mathbf{y}\in H(\Omega,\mathrm{div})$. This equation has a unique solution in $H(\Omega,\mathrm{div})$, since we can show that the left-hand side is a bounded and $H(\Omega,\mathrm{div})$-elliptic bilinear form and that the right-hand side is a linear form, thus we can apply the Lax-Milgram Theorem 2.2.

Assume that $\mathbf{y}\in \mathrm{span}\left\{\boldsymbol{\phi}^1,\boldsymbol{\phi}^2,\ldots,\boldsymbol{\phi}^M\right\} =: Y_h \subset H^1(\Omega,\mathbb{R}^2)$, i.e., $\mathbf{y} = \sum_{j=1}^{M} y_j\boldsymbol{\phi}^j$, with $M=2N$. This is no contradiction since $H^1\left(\Omega,\mathbb{R}^2\right)\subset H(\Omega,\mathrm{div})$. This leads to the linear system which we can solve

numerically:

$$\sum_{j=1}^{M} y_j \left(\beta \int_\Omega \left\langle \mathbf{A}^{-1}\boldsymbol{\phi}^j, \boldsymbol{\phi}^i \right\rangle \mathrm{d}\mathbf{x} + \frac{C_{F\Omega}^2}{\alpha^{\mathrm{ell}}}(1+\gamma) \int_\Omega \operatorname{div}\boldsymbol{\phi}^j \operatorname{div}\boldsymbol{\phi}^i \,\mathrm{d}\mathbf{x} - \frac{C_{T\Gamma_N}^2}{\alpha^{\mathrm{ell}}} \frac{1+\gamma}{\gamma} \int_{\Gamma_N} \left\langle \boldsymbol{\phi}^j, \mathbf{n} \right\rangle \left\langle \boldsymbol{\phi}^i, \mathbf{n} \right\rangle \mathrm{d}s \right)$$
$$= \beta \int_\Omega \left\langle \nabla v, \boldsymbol{\phi}^i \right\rangle \mathrm{d}\mathbf{x} + \frac{C_{F\Omega}^2}{\alpha^{\mathrm{ell}}}(1+\gamma) \int_\Omega \left(\langle \mathbf{b}, \nabla v \rangle + cv - f \right) \operatorname{div}\boldsymbol{\phi}^i \,\mathrm{d}\mathbf{x} - \frac{C_{T\Gamma_N}^2}{\alpha^{\mathrm{ell}}} \frac{1+\gamma}{\gamma} \int_{\Gamma_N} g_N \left\langle \boldsymbol{\phi}^i, \mathbf{n} \right\rangle \mathrm{d}s.$$

Denote the corresponding system matrices and vectors as follows

$$\begin{aligned}
(S_{i,j})_{i,j=1}^M &= \int_\Omega \operatorname{div}\boldsymbol{\phi}^j \operatorname{div}\boldsymbol{\phi}^i \,\mathrm{d}\mathbf{x}, & (z_i)_{i=1}^M &= \int_\Omega \left(\langle \mathbf{b}, \nabla v \rangle + cv - f \right) \operatorname{div}\boldsymbol{\phi}^i \,\mathrm{d}\mathbf{x}, \\
(K_{i,j})_{i,j=1}^M &= \int_\Omega \left\langle \mathbf{A}^{-1}\boldsymbol{\phi}^j, \boldsymbol{\phi}^i \right\rangle \mathrm{d}\mathbf{x}, & (g_i)_{i=1}^M &= \int_\Omega \left\langle \nabla v, \boldsymbol{\phi}^i \right\rangle \mathrm{d}\mathbf{x}, \\
(B_{i,j})_{i,j=1}^M &= \int_{\Gamma_N} \left\langle \boldsymbol{\phi}^j, \mathbf{n} \right\rangle \left\langle \boldsymbol{\phi}^i, \mathbf{n} \right\rangle \mathrm{d}s, & (k_i)_{i=1}^M &= \int_{\Gamma_N} g_N \left\langle \boldsymbol{\phi}^i, \mathbf{n} \right\rangle \mathrm{d}s.
\end{aligned} \tag{5.11}$$

We solve the following linear system for $\mathbf{y} = (y_j)_{j=1}^M$:

$$\left(\beta K + \frac{C_{F\Omega}^2}{\alpha^{\mathrm{ell}}}(1+\gamma) S - \frac{C_{T\Gamma_N}^2}{\alpha^{\mathrm{ell}}} \frac{1+\gamma}{\gamma} B \right) \mathbf{y} = \beta \mathbf{g} + \frac{C_{F\Omega}^2}{\alpha^{\mathrm{ell}}}(1+\gamma)\mathbf{z} - \frac{C_{T\Gamma_N}^2}{\alpha^{\mathrm{ell}}} \frac{1+\gamma}{\gamma} \mathbf{k}. \tag{5.12}$$

Then, we can compute the corresponding value of the majorant:

$$\begin{aligned}
\mathcal{M}_{\beta,\gamma}^2(v, \mathbf{y}) = (1+\beta) &\left(\mathbf{y}^\top K \mathbf{y} - 2\mathbf{y}^\top \mathbf{g} + \|\mathbf{A}\nabla v\|_{\mathbf{A}^{-1}}^2 \right) \\
&+ \frac{1+\beta}{\beta} \frac{C_{F\Omega}^2}{\alpha^{\mathrm{ell}}}(1+\gamma) \left(\mathbf{y}^\top S \mathbf{y} - 2\mathbf{y}^\top \mathbf{z} + \|f - \langle \mathbf{b}, \nabla v \rangle - cv\|_{L^2(\Omega)}^2 \right) \\
&+ \frac{1+\beta}{\beta} \frac{C_{T\Gamma_N}^2}{\alpha^{\mathrm{ell}}} \frac{1+\gamma}{\gamma} \left(\mathbf{y}^\top B \mathbf{y} - 2\mathbf{y}^\top \mathbf{k} + \|g_N\|_{L^2(\Gamma_N)}^2 \right).
\end{aligned} \tag{5.13}$$

We use an iteration procedure to minimize the majorant, stated in Algorithm 2. First, we use some gradient recovery procedure (see Section 2.5) of ∇v_h as initial guess $\mathbf{y}_0$. Then, we compute the three parts of the majorant and thus get starting values β_0 and γ_0, which corresponds in minimizing the majorant with respect to β and γ, respectively. Then, a first approximation of the majorant can be derived. For some (small) number of iteration steps or tolerance, this procedure is repeated, starting with solving (5.12) to get an approximation $\mathbf{y}_j$ and continuing with β_j and γ_j.
Another approach to minimize the majorant is stated in Algorithm 3. First, we assign γ and β with a certain value (e.g., 1). Then, the majorant is minimized with respect to $\mathbf{y}$ (which amounts to solving (5.12)). With this solution, we recompute the majorant and find a new γ and β by minimizing the majorant with respect to γ and β. The process is repeated for either some prescribed (small) amount of iteration steps or some (moderate) tolerance.

Algorithm 2 Minimization of the majorant

Input: v_h, ϕ_i, f, $\mathbf{A}$, $\mathbf{b}$, c, g_N, $C_{F\Omega}$, $C_{T\Gamma_N}$ and $I_{\max}$

Compute $\|f - \langle \mathbf{b}, \nabla v_h \rangle - c v_h\|^2_{L^2(\Omega)}$, $\|\mathbf{A}\nabla v_h\|^2_{\mathbf{A}^{-1}}$ and $\|g_N\|^2_{L^2(\Gamma_N)}$.

Assemble matrices S, K and B, and vectors $\mathbf{z}$, $\mathbf{g}$ and $\mathbf{k}$ as in (5.11).

Derive the recovered gradient $\mathbf{y}_0 = \mathbf{A}G_h(\nabla v_h)$.

Compute the three parts of the majorant (j=0):

$$\begin{aligned} \mathcal{M}^2_{\mathrm{Eq}} &= \mathbf{y}_j^\top S \mathbf{y}_j - 2\mathbf{y}_j^\top \mathbf{z} + \|f - \langle \mathbf{b}, \nabla v_h \rangle - c v_h\|^2_{L^2(\Omega)} \\ \mathcal{M}^2_{\mathrm{D}} &= \mathbf{y}_j^\top K \mathbf{y}_j - 2\mathbf{y}_j^\top \mathbf{g} + \|\mathbf{A}\nabla v_h\|^2_{\mathbf{A}^{-1}} \\ \mathcal{M}^2_{\Gamma} &= \mathbf{y}_j^\top B \mathbf{y}_j - 2\mathbf{y}_j^\top \mathbf{k} + \|g_N\|^2_{L^2(\Gamma_N)} \end{aligned} \tag{5.14}$$

Compute the value of β_0:

$$\beta_0 = \frac{C_{F\Omega}\mathcal{M}_{\mathrm{Eq}} + C_{T\Gamma_N}\mathcal{M}_{\Gamma}}{\sqrt{\alpha^{\mathrm{ell}}}\mathcal{M}_{\mathrm{D}}} \tag{5.15}$$

Compute the value of γ_0:

$$\gamma_0 = \frac{C_{T\Gamma_N}\mathcal{M}_{\Gamma}}{C_{F\Omega}\mathcal{M}_{\mathrm{Eq}}} \tag{5.16}$$

Compute the coarse upper bound of the error:

$$\mathcal{M}^2_{\beta_0,\gamma_0}(v_h, \mathbf{y}_0) = (1+\beta_0)\mathcal{M}^2_{\mathrm{D}} + \frac{1+\beta_0}{\beta_0}\frac{C^2_{F\Omega}}{\alpha^{\mathrm{ell}}}(1+\gamma_0)\mathcal{M}^2_{\mathrm{Eq}} + \frac{1+\beta_0}{\beta_0}\frac{C^2_{T\Gamma_N}}{\alpha^{\mathrm{ell}}}\frac{1+\gamma_0}{\gamma_0}\mathcal{M}^2_{\Gamma} \tag{5.17}$$

for $j = 1$ **to** $I_{\max}$ **do**

Solve the system:

$$\left(\beta_{j-1}K + \frac{C^2_{F\Omega}}{\alpha^{\mathrm{ell}}}(1+\gamma_{j-1})S - \frac{C^2_{T\Gamma_2}}{\alpha^{\mathrm{ell}}}\frac{1+\gamma_{j-1}}{\gamma_{j-1}}B\right)\mathbf{y}_j = \beta_{j-1}\mathbf{g} + \frac{C^2_{F\Omega}}{\alpha^{\mathrm{ell}}}(1+\gamma_{j-1})\mathbf{z} - \frac{C^2_{T\Gamma_N}}{\alpha^{\mathrm{ell}}}\frac{1+\gamma_{j-1}}{\gamma_{j-1}}\mathbf{k}$$

Compute $\mathcal{M}^2_{\mathrm{Eq}}$, $\mathcal{M}^2_{\mathrm{D}}$ and $\mathcal{M}^2_{\Gamma}$ by (5.14)

Compute β_j by (5.15)

Compute γ_j by (5.16)

Compute $\mathcal{M}^2_{\beta_j,\gamma_j}(v_h, \mathbf{y}_j)$ by (5.17)

end for

Output: $\mathcal{M}^2_{\beta_{I_{\max}},\gamma_{I_{\max}}}(v_h, \mathbf{y}_{I_{\max}})$ and $\mathbf{y} = \mathbf{y}_{I_{\max}}$

Algorithm 3 Minimization of the majorant, second approach

Input: v_h, $\boldsymbol{\phi}_i$, f, $\mathbf{A}$, $\mathbf{b}$, c, g_N, $C_{F\Omega}$, $C_{T\Gamma_N}$ and $I_{\max}$

Compute $\|f - \langle \mathbf{b}, \nabla v_h \rangle - c v_h\|^2_{L^2(\Omega)}$, $\|\mathbf{A}\nabla v_h\|^2_{\mathbf{A}^{-1}}$ and $\|g_N\|^2_{L^2(\Gamma_N)}$.

Assemble matrices S, K and B, and vectors $\mathbf{z}$, $\mathbf{g}$ and $\mathbf{k}$ as in (5.11).

Denote $\beta_0 = 1$ and $\gamma_0 = 1$.

for $j = 1$ **to** $I_{\max}$ **do**

Solve the system:

$$\left(\beta_{j-1}K + \frac{C^2_{F\Omega}}{\alpha^{\mathrm{ell}}}(1+\gamma_{j-1})S - \frac{C^2_{T\Gamma_N}}{\alpha^{\mathrm{ell}}}\frac{1+\gamma_{j-1}}{\gamma_{j-1}}B\right)\mathbf{y}_j = \beta_{j-1}\mathbf{g} + \frac{C^2_{F\Omega}}{\alpha^{\mathrm{ell}}}(1+\gamma_{j-1})\mathbf{z} - \frac{C^2_{T\Gamma_N}}{\alpha^{\mathrm{ell}}}\frac{1+\gamma_{j-1}}{\gamma_{j-1}}\mathbf{k}$$

Compute the three parts of the majorant:

$$\begin{aligned}
\mathcal{M}^2_{\mathrm{Eq}} &= \mathbf{y}_j^\top S\mathbf{y}_j - 2\mathbf{y}_j^\top\mathbf{z} + \|f - \langle \mathbf{b}, \nabla v_h\rangle - c v_h\|^2_{L^2(\Omega)}\\
\mathcal{M}^2_{\mathrm{D}} &= \mathbf{y}_j^\top K\mathbf{y}_j - 2\mathbf{y}_j^\top\mathbf{g} + \|\mathbf{A}\nabla v_h\|^2_{\mathbf{A}^{-1}}\\
\mathcal{M}^2_{\Gamma} &= \mathbf{y}_j^\top B\mathbf{y}_j - 2\mathbf{y}_j^\top\mathbf{k} + \|g_N\|^2_{L^2(\Gamma_N)}
\end{aligned}$$

Compute the new value of β:

$$\beta_j = \frac{C_{F\Omega}\mathcal{M}_{\mathrm{Eq}} + C_{T\Gamma_N}\mathcal{M}_{\Gamma}}{\sqrt{\alpha^{\mathrm{ell}}}\mathcal{M}_{\mathrm{D}}}$$

Compute the new value of γ:

$$\gamma_j = \frac{C_{T\Gamma_N}\mathcal{M}_{\Gamma}}{C_{F\Omega}\mathcal{M}_{\mathrm{Eq}}}$$

Compute the majorant:

$$\mathcal{M}^2_{\beta_j,\gamma_j}(v_h, \mathbf{y}_j) = (1+\beta_j)\mathcal{M}^2_{\mathrm{D}} + \frac{1+\beta_j}{\beta_j}\frac{C^2_{F\Omega}}{\alpha^{\mathrm{ell}}}(1+\gamma_j)\mathcal{M}^2_{\mathrm{Eq}} + \frac{1+\beta_j}{\beta_j}\frac{C^2_{T\Gamma_N}}{\alpha^{\mathrm{ell}}}\frac{1+\gamma_j}{\gamma_j}\mathcal{M}^2_{\Gamma}$$

end for

Output: $\mathcal{M}^2_{\beta_{I_{\max}},\gamma_{I_{\max}}}(v_h, \mathbf{y}_{I_{\max}})$ and $\mathbf{y} = \mathbf{y}_{I_{\max}}$

Remark 5.2. *In the above algorithms one has to choose $I_{\max}$ or some tolerance to stop the for-loop. We want to compute the majorant in an economic way, but since solving the mentioned system can be quite expensive, it is desirable to set $I_{\max}$ rather small or the tolerance only moderate. With the help of numerical experiments one can find typical numbers for $I_{\max}$ and the tolerance, which balance the accuracy and the computational cost. On the other hand, one could choose $\beta = \gamma = 1$, for simplicity, and hence minimize*

$$\mathcal{M}^2(v, \mathbf{y}) \le 2\|\mathbf{y} - \mathbf{A}\nabla v\|^2_{\mathbf{A}^{-1}} + 2\frac{1}{\alpha^{\mathrm{ell}}}\left(2C^2_{F\Omega}\|r_\Omega(v,\mathbf{y})\|^2_{L^2(\Omega)} + 2C^2_{T\Gamma_N}\|g_N - \langle \mathbf{y}, \mathbf{n}\rangle\|^2_{L^2(\Gamma_N)}\right)$$

with respect to $\mathbf{y}$. This would lead to Algorithm 2, without the steps of computing $\beta_0, \gamma_0, \beta_j$ and γ_j and replacing their values by one.

For the implementation, we first have to discuss how to calculate the additional matrices. They contain the divergence operator and for that we need vector valued basis functions. In our case, it is sufficient to take a combination of the scalar valued nodal basis functions. Since we want to keep the computational cost minimal for the majorant, we will always restrict to linear basis functions for the majorant and also for the gradient recovery. Thus, we consider the following vector valued basis functions

$$\left\{\widehat{\boldsymbol{\phi}}^i(\hat{\mathbf{x}}) \,\middle|\, i = 1,2,\dots,6\right\} = \left\{\begin{bmatrix}\widehat{\psi}_i(\hat{\mathbf{x}})\\ 0\end{bmatrix}, \begin{bmatrix}0\\ \widehat{\psi}_i(\hat{\mathbf{x}})\end{bmatrix} \,\middle|\, i = 1,2,3\right\},$$

with $\widehat{\psi}_i$ defined in (5.2). For the gradient, we than have to consider $\nabla_{\hat{\mathbf{x}}}\widehat{\phi}_1^i(\hat{\mathbf{x}})$ and $\nabla_{\hat{\mathbf{x}}}\widehat{\phi}_2^i(\hat{\mathbf{x}})$.
For the majorant we made the ansatz $\mathbf{y} = \sum_{i=1}^{M} y_i \boldsymbol{\phi}^i$ and we have to compute the matrices and vectors K, S, $\mathbf{g}$ and $\mathbf{z}$, defined in (5.11). Further, for Neumann boundary conditions we need the matrix B and the vector $\mathbf{k}$. We will explain in the following how to compute matrix S, which is similar as deriving $a(\psi_j, \psi_i)$:

$$S_{i,j} = \int_\Omega \operatorname{div} \boldsymbol{\phi}^j \operatorname{div} \boldsymbol{\phi}^i \,\mathrm{d}\mathbf{x} = \sum_{T\in\mathcal{T}_h} \int_T \operatorname{div} \boldsymbol{\phi}^j \operatorname{div} \boldsymbol{\phi}^i \,\mathrm{d}\mathbf{x} =: \sum_{T\in\mathcal{T}_h} S_{T;i,j}.$$

Again, we transform the integral onto T_{ref}:

$$S_{T;i,j} = |2|T|| \int_{T_{\mathrm{ref}}} \left(\sum_{m=1}^{2} \left\langle \nabla_{\hat{\mathbf{x}}} \widehat{\phi}_m^j(\hat{\mathbf{x}}), (\mathrm{D}\chi_T)^{-1} \mathbf{e}_m \right\rangle \right) \left(\sum_{m=1}^{2} \left\langle \nabla_{\hat{\mathbf{x}}} \widehat{\phi}_m^i(\hat{\mathbf{x}}), (\mathrm{D}\chi_T)^{-1} \mathbf{e}_m \right\rangle \right) \mathrm{d}\hat{\mathbf{x}},$$

where we used the relation

$$\operatorname{div} \widehat{\boldsymbol{\phi}} = \left\langle \nabla_{\hat{\mathbf{x}}} \widehat{\phi}_1, (\mathrm{D}\chi_T)^{-1} \mathbf{e}_1 \right\rangle + \left\langle \nabla_{\hat{\mathbf{x}}} \widehat{\phi}_2, (\mathrm{D}\chi_T)^{-1} \mathbf{e}_2 \right\rangle, \tag{5.18}$$

with $\mathbf{e}_1 = [1,0]^\top$ and $\mathbf{e}_2 = [0,1]^\top$. Further, we use again simplex coordinates

$$\begin{aligned} S_{T;i,j} = |2|T|| \int_0^1 \int_0^1 \xi \Bigg(& \sum_{m=1}^{2} \left\langle (\mathrm{D}\chi_T)^{-\top} (\mathrm{D}F)^{-\top} \nabla_{(\xi,\eta)} \widehat{\phi}_m^j(\xi, \xi\eta), \mathbf{e}_m \right\rangle \Bigg) \\ & \Bigg(\sum_{m=1}^{2} \left\langle (\mathrm{D}\chi_T)^{-\top} (\mathrm{D}F)^{-\top} \nabla_{(\xi,\eta)} \widehat{\phi}_m^i(\xi, \xi\eta), \mathbf{e}_m \right\rangle \Bigg) \,\mathrm{d}\xi \,\mathrm{d}\eta \end{aligned}$$

and apply a Gauss quadrature, hence we compute:

$$\begin{aligned} S_{T;i,j} = |2|T|| \sum_{k=1}^{n_Q} \sum_{l=1}^{n_Q} & w_{k,n_Q} w_{l,n_Q} x_{k,n_Q} \\ & \left(\sum_{m=1}^{2} \left\langle (\mathrm{D}\chi_T)^{-\top} (\mathrm{D}F)_{k,l}^{-\top} \nabla_{(x_{k,n_Q}, x_{l,n_Q})} \widehat{\phi}_m^j(x_{k,n_Q}, x_{k,n_Q} x_{l,n_Q}), \mathbf{e}_m \right\rangle \right) \\ & \left(\sum_{m=1}^{2} \left\langle (\mathrm{D}\chi_T)^{-\top} (\mathrm{D}F)_{k,l}^{-\top} \nabla_{(x_{k,n_Q}, x_{l,n_Q})} \widehat{\phi}_m^i(x_{k,n_Q}, x_{k,n_Q} x_{l,n_Q}), \mathbf{e}_m \right\rangle \right), \end{aligned}$$

with $(\mathrm{D}\chi_T)^{-\top}$ and $(\mathrm{D}F)_{k,l}^{-\top}$ defined as before.
The matrix K can be computed similar to part 3, but with a scalar product as in part 1. The vectors $\mathbf{g}$ and $\mathbf{z}$ are also easily computed with the considerations from before, where we get again a factor ± 1 for $\mathbf{z}$, the only new term is

$$\nabla v \left(\chi_T(x_{k,n_Q}, x_{k,n_Q} x_{l,n_Q}) \right).$$

This can be treated in the following way: The values of v are given in the nodes $\mathbf{x}_i$, so we use interpolation with basis functions $\nabla\psi_i$.
For the evaluation of the norms $\|f - \langle \mathbf{b}, \nabla v \rangle - cv\|_{L^2(\Omega)}^2$ and $\|\mathbf{A}\nabla v\|_{\mathbf{A}^{-1}}^2$, we further need

$$v \left(\chi_T(x_{k,n_Q}, x_{k,n_Q} x_{l,n_Q}) \right),$$

which can be evaluated as an interpolation with ψ_i.
For B, $\mathbf{k}$ and $\|g_N\|_{L^2(\Gamma_N)}^2$, we consider again an integration over an edge. The normal vector $\mathbf{n}$ on the edge $[\mathbf{x}_B, \mathbf{x}_C]$, for a general triangle with vertices $\mathbf{x}_A$, $\mathbf{x}_B$ and $\mathbf{x}_C$, can be derived as

$$\mathbf{n} = \frac{1}{\|\mathbf{x}_C - \mathbf{x}_B\|_2} \begin{bmatrix} y_C - y_B \\ x_B - x_C \end{bmatrix}. \tag{5.19}$$

The outer normal vector is then derived by multiplying with the sign of the area of the triangle.
The computation of the cell majorant from Proposition 4.5 is very similar, we only have to change the matrices S and K and the vectors $\mathbf{z}$ and $\mathbf{g}$ slightly.

5.2.6 Gradient Recovery

Consider the gradient recovery procedure as explained in Section 2.5, i.e.,

$$G_h : L^2\left(\Omega, \mathbb{R}^2\right) \to H^1\left(\Omega, \mathbb{R}^2\right),$$

with $G_h\left(\nabla v_h\right) := S^m Q_h \nabla v_h$. We will describe in the following how one can compute the L^2-projection and the smoothing operator S, in order to derive the gradient recovery.
Recall the definition of the discrete L^2-projection operator: $Q_h : L^2\left(\Omega, \mathbb{R}^2\right) \to H^1\left(\Omega, \mathbb{R}^2\right)$ defined by

$$\left(Q_h \nabla v_h, \mathbf{w}_h\right)_{L^2(\Omega)} = \left(\nabla v_h, \mathbf{w}_h\right)_{L^2(\Omega)}, \quad \forall \mathbf{w}_h \in L^2\left(\Omega, \mathbb{R}^2\right). \tag{5.20}$$

We consider as before the vector valued linear basis functions

$$\left\{\widehat{\boldsymbol{\phi}}^i(\hat{\mathbf{x}}) \,\middle|\, i = 1, 2, \ldots, 6\right\} = \left\{\begin{bmatrix}\widehat{\psi}_i(\hat{\mathbf{x}}) \\ 0\end{bmatrix}, \begin{bmatrix}0 \\ \widehat{\psi}_i(\hat{\mathbf{x}})\end{bmatrix} \,\middle|\, i = 1, 2, 3\right\},$$

with $\widehat{\psi}_i$ defined in (5.2). The right-hand side of (5.20) can be written as

$$\left(\nabla v_h, \boldsymbol{\phi}^i\right)_{L^2(\Omega)} = \sum_{j=1}^{N} v_j \left(\nabla \psi_j, \boldsymbol{\phi}^i\right)_{L^2(\Omega)}, \quad \forall i = 1, \ldots, M,$$

where

$$\left(\nabla \psi_j, \boldsymbol{\phi}^i\right)_{L^2(\Omega)} = \begin{cases} \left(\partial_x \psi_j, \psi_i\right)_{L^2(\Omega)} & i = 1, \ldots, N, \\ \left(\partial_y \psi_j, \psi_{i-N}\right)_{L^2(\Omega)} & i = N+1, \ldots, M. \end{cases}$$

Defining matrices $\mathbf{B}_x$ and $\mathbf{B}_y$ with the anti-symmetric part of the system matrix for vectors $\mathbf{b} = [1, 0]^\top$ and $\mathbf{b} = [0, 1]^\top$, respectively, we get

$$\begin{aligned} \left(\mathbf{B}_x\right)_{i,j} &= \int_\Omega \left\langle [1,0]^\top, \nabla \psi_j \right\rangle \psi_i \, \mathrm{d}\mathbf{x} &&= \int_\Omega \left(\partial_x \psi_j\right) \psi_i \, \mathrm{d}\mathbf{x} \\ \left(\mathbf{B}_y\right)_{i-N,j} &= \int_\Omega \left\langle [0,1]^\top, \nabla \psi_j \right\rangle \psi_{i-N} \, \mathrm{d}\mathbf{x} &&= \int_\Omega \left(\partial_y \psi_j\right) \psi_i \, \mathrm{d}\mathbf{x}. \end{aligned}$$

Hence, we have

$$\left(\nabla v_h, \boldsymbol{\phi}^i\right)_{L^2(\Omega)} = \sum_{j=1}^{N} v_j \begin{cases} \mathbf{B}_x & i = 1, \ldots, N, \\ \mathbf{B}_y & i = N+1, \ldots, M. \end{cases} \tag{5.21}$$

Further, we can write $Q_h \nabla v_h$ by using the basis functions:

$$Q_h \nabla v_h = \sum_{j=1}^{M} q_j \boldsymbol{\phi}^j = \sum_{j=1}^{N} q_j \begin{bmatrix}\psi_j \\ 0\end{bmatrix} + \sum_{j=N+1}^{M} q_j \begin{bmatrix}0 \\ \psi_{j-N}\end{bmatrix} = \sum_{j=1}^{N} \begin{bmatrix}q_j \\ q_{j+N}\end{bmatrix} \psi_j,$$

for some constant coefficients q_j for $j = 1, 2, \ldots, M$. Then, the left-hand side of (5.20) takes the form:

$$\begin{aligned} \left(Q_h \nabla v_h, \boldsymbol{\phi}^i\right)_{L^2(\Omega)} &= \sum_{j=1}^{N} \left(\begin{bmatrix}q_j \\ q_{j+N}\end{bmatrix} \psi_j, \boldsymbol{\phi}^i \right)_{L^2(\Omega)} \\ &= \begin{cases} \sum_{j=1}^{N} q_j \left(\psi_j, \psi_i\right)_{L^2(\Omega)} & i = 1, \ldots, N, \\ \sum_{j=1}^{N} q_{j+N} \left(\psi_j, \psi_{i-N}\right)_{L^2(\Omega)} & i = N+1, \ldots, M. \end{cases} \end{aligned} \tag{5.22}$$

Note that the matrix $\mathbf{M}$, where $\mathbf{M}_{i,j} = \left(\psi_j, \psi_i\right)_{L^2(\Omega)}$, is exactly part 3 of the system matrix in (5.4) for $c = 1$ and usually called mass matrix. Finally, by putting together (5.21) and (5.22), we get two decoupled equations

$$\mathbf{M}\mathbf{q}_x = \mathbf{B}_x \mathbf{v}, \quad \mathbf{M}\mathbf{q}_y = \mathbf{B}_y \mathbf{v}, \tag{5.23}$$

where $\mathbf{v} = \left(v_j\right)_{j=1}^N$, $\mathbf{q}_x = \left(q_j\right)_{j=1}^N$ and $\mathbf{q}_y = \left(q_j\right)_{j=N+1}^M$. Solving these equations we get the coefficients $\mathbf{q}_x$ and $\mathbf{q}_y$, hence we can compute the L^2-projection. The effect of the L^2-projection can be illustrated in a one dimensional example, see Figure 5.7.

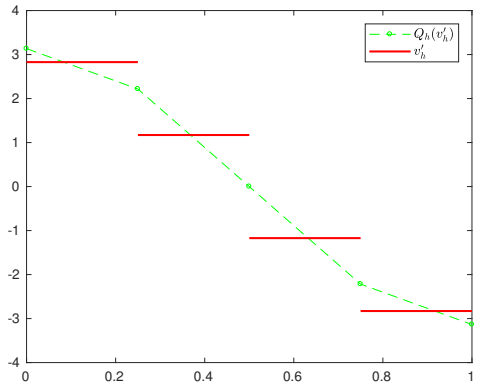


Figure 5.7: Gradient recovery $Q_h \nabla v_h$ compared to the piecewise constant ∇v_h.

Recall the construction of the smoothing operator S:

$$S = I - \lambda^{-1} A_h,$$

where I is the identity operator, $\lambda = \rho(A_h) \leq c_{\text{inv}}^2 h^{-2}$ and A_h the discrete operator defined by

$$(A_h u_h, v_h)_{L^2(\Omega)} := (\nabla u_h, \nabla v_h)_{L^2(\Omega)} + (u_h, v_h)_{L^2(\Omega)}, \quad \forall u_h, v_h \in V_h.$$

We use the basis functions to get the discrete form of the operator: Let

$$S v_h = \sum_{j=1}^{N} s_j \psi_j,$$

for some constant vector $\mathbf{s} = (s_j)_{j=1}^N$. Then, by taking the L^2-scalar product with a test function ψ_i, we get

$$(S v_h, \psi_i)_{L^2(\Omega)} = \sum_{j=1}^{N} s_j (\psi_j, \psi_i)_{L^2(\Omega)} = \mathbf{M}\mathbf{s}.$$

Doing the same for the operator definition, we have

$$\left(\left(I - \lambda^{-1} A_h\right) v_h, \psi_i\right)_{L^2(\Omega)} = \sum_j v_j \left((\psi_j, \psi_i)_{L^2(\Omega)} - \lambda^{-1} (A_h \psi_j, \psi_i)_{L^2(\Omega)}\right) = \left(\mathbf{M} - \lambda^{-1} \mathbf{A}_h\right) \mathbf{v},$$

where $\mathbf{A}_h$ is the system matrix, which is defined over (5.4) for $\mathbf{A} = I$, $\mathbf{b} = \mathbf{0}$ and $c = 1$. In conclusion, we get

$$\mathbf{s} = \left(\mathbf{I} - \lambda^{-1} \mathbf{M}^{-1} \mathbf{A}_h\right) \mathbf{v}.$$

Hence, for the gradient recovery procedure we multiply the vectors $\mathbf{q}_x$ and $\mathbf{q}_y$ separately m-times with the matrix

$$\mathbf{S}_h = \mathbf{I} - \lambda^{-1} \mathbf{M}^{-1} \mathbf{A}_h. \tag{5.24}$$

Note that we can neglect $\mathbf{M}^{-1}$ in the sense, that we include it into the scaling factor λ^{-1}, i.e., we denote

$$\mathbf{S}_h = \mathbf{I} - \widetilde{\lambda}^{-1} \mathbf{A}_h$$

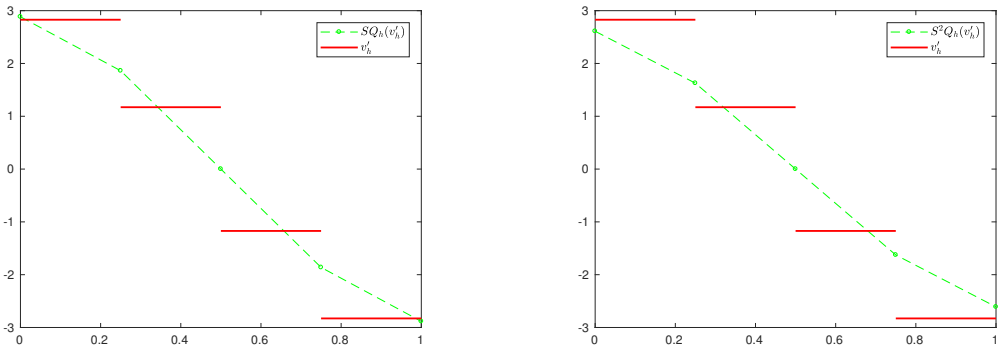
and consider $\widetilde{\lambda} = \rho(\mathbf{A}_h)$. Then, it holds $\widetilde{\lambda} = O(1)$, which is due to the fact that the eigenvalues of the mass matrix $\mathbf{M}$ are of order $O\left(h^{-2}\right)$, depending on the uniformity constant, see e.g. [16]. Hence, the factor $\widetilde{\lambda}$ should be chosen such that the spectrum of the smoothing matrix $\mathbf{S}_h$ is between 0 and 1. This is due to the connection to iterative solvers for linear systems, where the iteration matrix should have spectral radius smaller than 1.
There are several possibilities for the choice of an upper bound for the scaling factor $\widetilde{\lambda}$:

First, we could decide to not scale the matrix at all, hence $\mathbf{S}_1 = (\mathbf{I} - \mathbf{A}_h)$, which typically does not give a spectral radius smaller than 1.

Secondly, we can take the approximate maximal eigenvalue of $\mathbf{A}_h$ as scaling factor, hence $\mathbf{S}_2 = (\mathbf{I} - \widetilde{\lambda}^{-1}\mathbf{A}_h)$. For this we can use $\widetilde{\lambda} \leq \|\mathbf{A}_h\|_\infty$, since $\mathbf{A}_h$ is symmetric.

Finally, we can scale the matrix $\mathbf{A}_h$ by the absolute row sum. For this consider the matrix $\mathbf{D} = (d_{i,j})_{i,j=1}^N$, where $d_{i,j} = 0$ for $i \neq j$ and $d_{i,i} = \sum_{j=1}^{N} |a_{i,j}|$ for $i = 1, \dots, N$, hence we get $\mathbf{S}_3 = (\mathbf{I} - \mathbf{D}^{-1}\mathbf{A}_h)$.

For a comparison of the different variants and the behaviour of the gradient recovery see Chapter 6. The effect of the smoothing operator, where we chose $\mathbf{S}_2$, can be illustrated in a one dimensional example, see Figure 5.8.



(a) $m = 1$ (b) $m = 2$

Figure 5.8: Gradient recovery $S^m Q_h \nabla v_h$ compared to the piecewise constant ∇v_h.

5.3 Two Scale Approximation

As explained in Chapter 4, we compute a two scale approximation $\widetilde{w}_{1,\varepsilon}^{(l,j)}$ defined in (4.10), which is composed of $u_0^{(l,j)}$, $\widehat{\mathbf{N}}^{(l)}$, φ_ε and $\mathbf{C}_h$. For that, we start with a coarse mesh $\mathcal{T}_{h_\varepsilon} = \bigcup_{\mathbf{i}} \mathcal{T}_{h,\mathbf{i}}^{\varepsilon}$ of Ω, depicted in Figure 5.9 top left, where $\mathcal{T}_{h,\mathbf{i}}^{\varepsilon} := \mathbf{x}_\mathbf{i} + \varepsilon\widehat{\mathcal{T}}_h$. Hence, for the initial mesh of Ω we have a macro mesh size $h_\varepsilon \leq \varepsilon$ and for the initial mesh of the cell we have a micro mesh size $h = \frac{h_\varepsilon}{\varepsilon} \leq 1$. On $\widehat{\mathcal{T}}_h$, depicted in Figure 5.9 bottom left, we compute first coarse approximations $\widehat{\mathbf{N}}^{(1)}$ and refine $l-1$ times to get $\widehat{\mathbf{N}}^{(l)}$. Then, we compute the approximation $\mathbf{A}_{0,l}$. Additional to that, we have a mesh $\mathcal{T}_H$ of Ω, depicted in Figure 5.9 top right, which does not have to coincide with $\mathcal{T}_{h_\varepsilon}$. Therefore, we further have a macro mesh size H. On $\mathcal{T}_H$ we compute a first approximation $u_0^{(l,1)}$ and refine $j-1$ times to get $u_0^{(l,j)}$.

Before we can continue, we have to apply the Clément operator to $\nabla u_0^{(l,j)}$. The Clément operator is explained in Section A.4 and can be implemented as follows, see also [17]: Similar to the gradient recovery procedure, we apply the Clément operator $\mathbf{C}_h$ to $\nabla v_h \in L^2(\Omega, \mathbb{R}^2)$, where $\nabla v_h = \sum_{i=1}^N v_j \nabla \psi_j$. Thus, we want to compute the vector valued coefficients $\boldsymbol{\gamma}_j, j = 1, \dots, N$, such that

$$\mathbf{C}_h \nabla v_h = \sum_{j=1}^{N} \boldsymbol{\gamma}_j \psi_j.$$

The difference is, that we consider now a local procedure. The vector $\boldsymbol{\gamma}_j$ is derived by a local patch

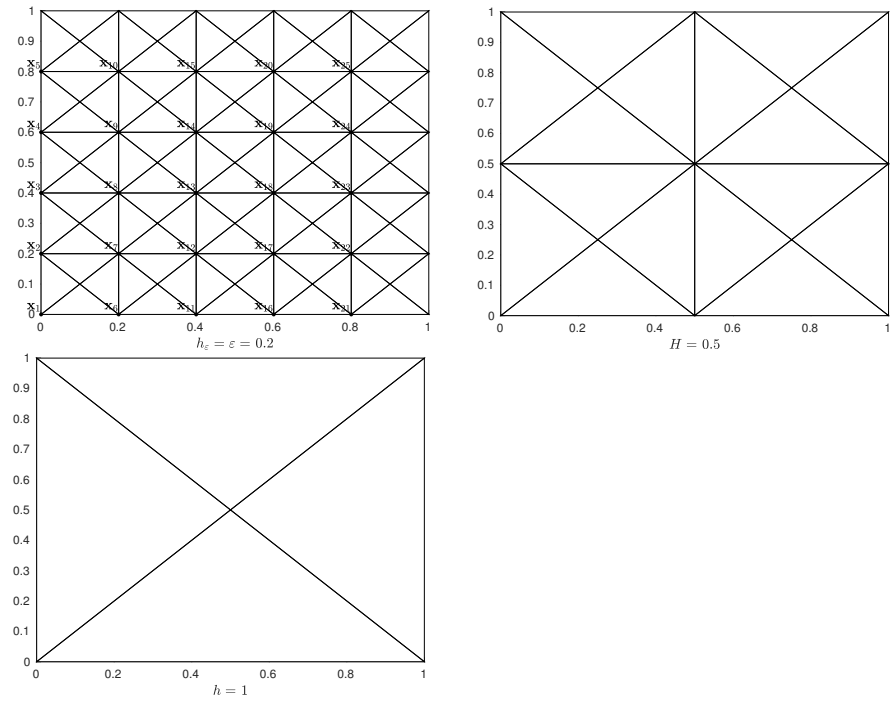


Figure 5.9: Initial mesh for the two scale approximation, the homogenized problem and the cell problem.

averaging, where we restrict ourselves to the case of linear basis functions:

$$\begin{aligned}\boldsymbol{\gamma}_j &= \frac{1}{|\omega_j|}\int_{\omega_j} \nabla v_h|_{\omega_j} \\ &= \frac{1}{|\omega_j|}\sum_{T\in\omega_j}\int_T \nabla v_h|_T \\ &= \frac{1}{|\omega_j|}\sum_{T\in\omega_j} |2|T||\sum_{i=1}^{3}\hat{v}_i \int_{T_{\text{ref}}} \left(\mathrm{D}\chi_T\right)^{-\top}\nabla_{\hat{\mathbf{x}}}\widehat{\psi}_j, \quad \forall i = 1,\dots,N,\end{aligned}$$

where $\hat{v}_i$ is equal to v_i, but with local numbering instead of global. The simplex coordinates and the quadrature rule are applied as in the sections before.
Then, for every $\mathbf{i}$, we compute

$$\widetilde{w}_{1,\varepsilon}^{(l,j)}(\mathbf{x}) := u_0^{(l,j)}(\mathbf{x}) - \varepsilon\varphi_\varepsilon(\mathbf{x})\left\langle \widehat{\mathbf{N}}^{(l)}\left(\frac{\mathbf{x}-\mathbf{x_i}}{\varepsilon}\right), \mathbf{C}_h\nabla u_0^{(l,j)}(\mathbf{x})\right\rangle, \quad \forall \mathbf{x}\in\mathcal{T}_{h,\mathbf{i}}^{\varepsilon}.$$

Since we have a different mesh for $u_0^{(l,j)}$, $\mathbf{x}$ does not have to be a node in $\mathcal{T}_H$ and thus we have to consider how to compute $u_0^{(l,j)}(\mathbf{x})$ and $\mathbf{C}_h\nabla u_0^{(l,j)}(\mathbf{x})$. For every node $\mathbf{x}\in\mathcal{T}_{h,\mathbf{i}}^{\varepsilon}$, we seek one triangle $T\in\mathcal{T}_H$ which contains $\mathbf{x}$. Then, by interpolation, we get the desired values in $\mathbf{x}$:

$$\begin{aligned}u_0^{(l,j)}(\mathbf{x}) &= \sum_{i=1}^{s}\left(u_0^{(l,j)}\right)_i \psi_i(\mathbf{x}), \\ \mathbf{C}_h\nabla u_0^{(l,j)}(\mathbf{x}) &= \sum_{i=1}^{s}\left(\boldsymbol{\gamma}^{(l,j)}\right)_i \psi_i(\mathbf{x}),\end{aligned}$$

where $\left(u_0^{(l,j)}\right)_i$ is the value of the discrete solution $u_0^{(l,j)}$ at the i-th node of T, with local degree of freedom s. The basis function then gets evaluated as $\psi_i(\mathbf{x}) = \widehat{\psi}_i\left(\chi_T^{-1}(\mathbf{x})\right)$.

In order to achieve the a priori convergence rates, we also compute the gradient of $\widetilde{w}_{1,\varepsilon}^{(l,j)}$ as a composition:

$$\begin{aligned}\nabla_{\mathbf{x}}\widetilde{w}_{1,\varepsilon}^{(l,j)} := \nabla_{\mathbf{x}}u_0^{(l,j)} - \varepsilon\nabla_{\mathbf{x}}\varphi_\varepsilon\left\langle\widehat{\mathbf{N}}^{(l)}, \mathbf{C}_h\nabla_{\mathbf{x}}u_0^{(l,j)}\right\rangle \\ - \varphi_\varepsilon\nabla_{\mathbf{y}}\widehat{\mathbf{N}}^{(l)}\left(\mathbf{C}_h\nabla_{\mathbf{x}}u_0^{(l,j)}\right) - \varepsilon\varphi_\varepsilon\nabla_{\mathbf{x}}\left(\mathbf{C}_h\nabla_{\mathbf{x}}u_0^{(l,j)}\right)\widehat{\mathbf{N}}^{(l)}.\end{aligned}$$

The cutoff function φ_ε, which corrects the boundary condition, could be chosen as

$$\varphi_\varepsilon(\mathbf{x}) = \min\left(1, \frac{\operatorname{dist}(\mathbf{x}, \partial\Omega)}{\varepsilon}\right),$$

which was already mentioned in (3.30). For the implementation we use a similar function, which again satisfies (3.29), illustrated in Figure 5.10.

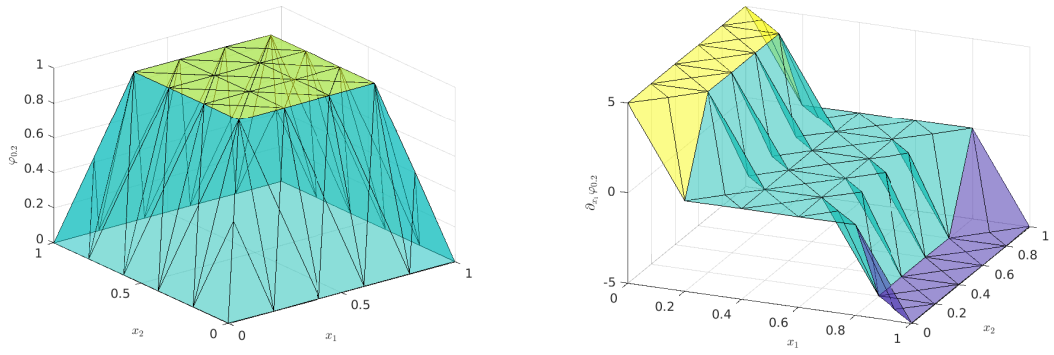


Figure 5.10: $\varphi_\varepsilon(\mathbf{x})$ and $\partial_{x_1}\varphi_\varepsilon(\mathbf{x})$ for $\varepsilon = 0.2$.

In Chapter 6 we will show the behaviour of the two scale approximation for different values of ε, h and H. If the value ε is not small enough, then we know a priori, that the two scale approximation will not be very accurate. Furthermore, h and H have to be sufficiently small, so that the discrete solutions of the cell and homogenized problems are of a desired accuracy. It will be interesting to observe the dependence of the three parameters on each other.

6 Numerical Experiments

In this chapter we present the results from our numerical experiments. In Section 6.1 we examine the performance of the gradient recovery procedure for different parameters. In Section 6.2 we study the behaviour of the majorant for a simple Dirichlet problem for different choices of parameters and algorithms. We conclude in Section 6.3 with three homogenization problems and analysing the sharpness of the majorants and the total error majorant described in Chapter 4.

6.1 Gradient Recovery

In the following we will present the superconvergence of the gradient recovery procedure explained in Section 2.5. The operator was defined by $G_h(\nabla v_h) = S^m Q_h \nabla v_h$ and thus we can choose the value $m \geq 0$ and the scaling factor λ, as described in Subsection 5.2.6.
We consider the Dirichlet problem

$$\begin{aligned} -\operatorname{div}(\nabla u) &= f \quad \text{in } \Omega = (0,1)^2, \\ u &= g \quad \text{on } \Gamma, \end{aligned} \tag{6.1}$$

with a smooth solution $u(\mathbf{x}) = e^{x_1+x_2}$ depicted in Figure 6.1, hence $f(\mathbf{x}) = -2e^{x_1+x_2}$ and $g(\mathbf{x}) = e^{x_1+x_2}$. Therefore, the regularity assumptions for the superconvergence argument are fulfilled.

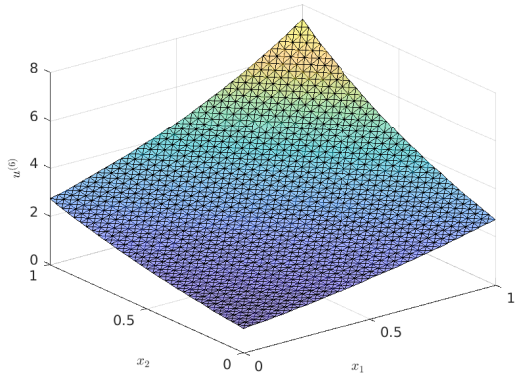


Figure 6.1: Approximated solution of problem (6.1).

Note that we consider a different initial mesh and a different refinement strategy than in [5] and [6], furthermore we only consider uniform refinement. Thus, the $O(h^{2\sigma})$ irregularity of the triangulation is not satisfied. Nevertheless, we observe superconvergence, as shown in the following, for different choices of parameters.
Consider the notation of the H^1- and L^2-norm, the error and the efficiency index of the gradient recovery:

$$\begin{aligned} H1 &:= \|\nabla(u - u_h)\|_{L^2} & L2 &:= \|u - u_h\|_{L^2} \\ Err_G &:= \|\nabla u - G_h(\nabla u_h)\|_{L^2} & \mathrm{Eff}_G &:= \frac{\|\nabla u_h - G_h(\nabla u_h)\|_{L^2}}{\|\nabla(u - u_h)\|_{L^2}} \end{aligned}$$

To compare the order of convergence p, we compute a least square fit of the data to a function of the form $F(N) = CN^{-p}$, as suggested in [6].

6.1.1 Comparison of the Scaling Factors

As explained in Subsection 5.2.6, we can either choose $\mathbf{S}_1 = (\mathbf{I} - \mathbf{A}_h)$, $\mathbf{S}_2 = (\mathbf{I} - \widetilde{\lambda}^{-1}\mathbf{A}_h)$, with $\widetilde{\lambda}$ approximated from above by $\|\mathbf{A}_h\|_\infty$, or $\mathbf{S}_3 = (\mathbf{I} - \mathbf{D}^{-1}\mathbf{A}_h)$, i.e. scaling with the absolute row sum, as the smoothing matrix.
In Table 6.1, the errors for those three choices are listed for the Dirichlet problem (6.1) and depending on the number of triangles nt. Further, the similar but different behaviour is illustrated in Figure 6.2.

			$\mathbf{S}_1$		$\mathbf{S}_2$		$\mathbf{S}_3$	
nt	$H1$	$L2$	Err_G	Eff_G	Err_G	Eff_G	Err_G	Eff_G
4	1.25e+0	3.85e-1	1.95e+0	1.70	6.98e-1	0.84	1.10e+0	1.04
16	6.04e-1	9.53e-2	1.28e+0	2.27	2.59e-1	0.98	4.20e-1	1.10
64	2.99e-1	2.38e-2	5.09e-1	1.92	9.80e-2	1.00	1.54e-1	1.06
256	1.49e-1	5.95e-3	1.93e-1	1.59	3.63e-2	1.00	5.55e-2	1.03
1024	7.44e-2	1.49e-3	7.06e-2	1.35	1.32e-2	1.00	1.98e-2	1.02
4096	3.72e-2	3.72e-4	2.54e-2	1.19	4.72e-3	1.00	7.03e-3	1.01
16384	1.86e-2	9.30e-5	9.06e-3	1.10	1.68e-3	1.00	2.49e-3	1.00
Order	1.01	2.00	1.34		1.45		1.47	

Table 6.1: Comparison of the error estimates for different scaling factors.

The order of convergence shows the linear and quadratic convergence of the H^1- and L^2-error, respectively, consistent with the a priori theory. The smoothing matrix $\mathbf{S}_1$ improves the convergence to some extent, but $\mathbf{S}_2$ and $\mathbf{S}_3$ give better results. Taking into account the efficiency index, we conclude in choosing $\mathbf{S}_2$ as the optimal smoothing matrix for this example.

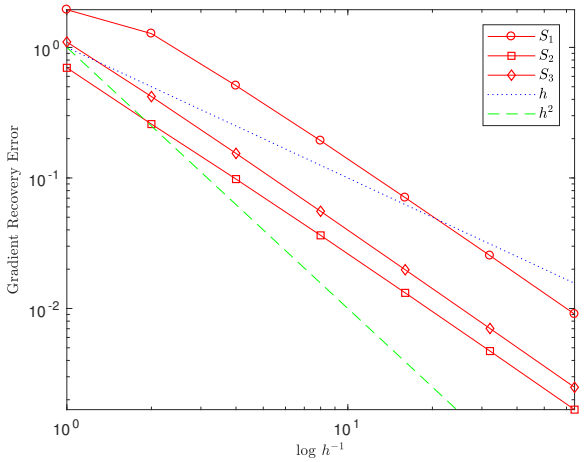


Figure 6.2: Convergence of the gradient recovery for different scaling factors.

6.1.2 Comparison of the Number of Smoothing Steps

For the gradient recovery operator $G_h(\nabla v_h) = S^m Q_h \nabla v_h$, we will choose from now on $S = \mathbf{S}_2$. Below, we vary the number of smoothing steps $m \geq 0$.

In Table 6.2, the errors for different values of m are listed, again for the Dirichlet problem (6.1). The effect of the number of smoothing steps is further illustrated in Figure 6.3.

			$m=0$		$m=1$		$m=2$	
nt	$H1$	$L2$	Err_G	Eff_G	Err_G	Eff_G	Err_G	Eff_G
4	1.25e+0	3.85e-1	6.69e-1	0.81	6.98e-1	0.84	7.81e-1	0.89
16	6.04e-1	9.53e-2	1.98e-1	0.94	2.59e-1	0.98	3.57e-1	1.05
64	2.99e-1	2.38e-2	6.58e-2	0.98	9.80e-2	1.00	1.42e-1	1.05
256	1.49e-1	5.95e-3	2.26e-2	0.99	3.63e-2	1.00	5.40e-2	1.03
1024	7.44e-2	1.49e-3	7.88e-3	0.99	1.32e-2	1.00	1.98e-2	1.02
4096	3.72e-2	3.72e-4	2.77e-3	1.00	4.72e-3	1.00	7.14e-3	1.01
16384	1.86e-2	9.30e-5	9.76e-4	1.00	1.68e-3	1.00	2.55e-3	1.00
Order	1.01	2.00	1.56		1.45		1.39	

	$m=3$		$m=10$	
nt	Err_G	Eff_G	Err_G	Eff_G
4	8.86e-1	0.95	1.56e+0	1.45
16	4.58e-1	1.15	1.02e+0	1.88
64	1.87e-1	1.12	4.70e-1	1.81
256	7.16e-2	1.07	1.83e-1	1.54
1024	2.64e-2	1.04	6.79e-2	1.32
4096	9.53e-3	1.02	2.46e-2	1.18
16384	3.40e-3	1.01	8.80e-3	1.10
Order	1.36		1.28	

Table 6.2: Error estimates as a function of m for uniform meshes.

The order of convergence shows that by only using the L^2-projection, i.e. $m = 0$, we achieve the best result. According to [6], we should see a dramatic improvement of the superconvergence for $m = 1, 2$. For this example and with our refinement strategy, we do not see this effect. Still, for $m = 1$, we observe superconvergence and the efficiency index appears to be very good. For $m \geq 2$, we see that the results deteriorate, but this parameter test should be performed for every problem separately, in order to get the best post-processing procedure according to the problem.

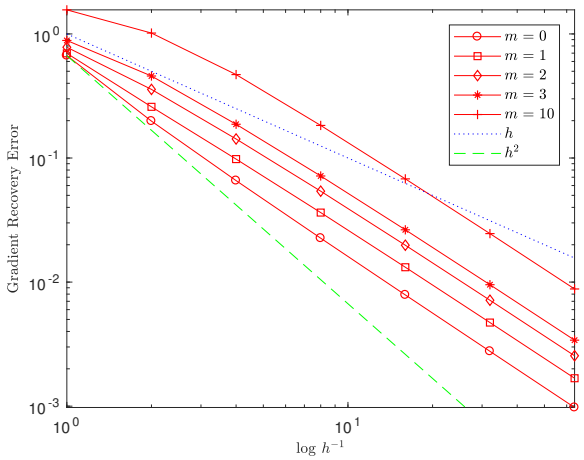


Figure 6.3: Convergence of the gradient recovery for different values m.

6.2 Majorant for the Dirichlet Problem

We consider two Dirichlet problems, the one from before

$$\begin{aligned} -\operatorname{div}(\nabla u) &= f \quad \text{in } \Omega = (0,1)^2, \\ u &= g \quad \text{on } \Gamma, \end{aligned} \tag{6.2}$$

with a smooth solution $u(\mathbf{x}) = e^{x_1+x_2}$, hence $f(\mathbf{x}) = -2e^{x_1+x_2}$ and $g(\mathbf{x}) = e^{x_1+x_2}$, and the same problem with diffusion matrix depending on $\mathbf{x}$

$$\begin{aligned} -\operatorname{div}(\mathbf{A}\nabla u) &= f \quad \text{in } \Omega = (0,1)^2, \\ u &= g \quad \text{on } \Gamma, \end{aligned} \tag{6.3}$$

with

$$\mathbf{A}(\mathbf{x}) := \begin{pmatrix} e^{x_1}e^{x_2} & 0 \\ 0 & e^{x_1}e^{x_2} \end{pmatrix},$$

where u and g are unchanged and $f(\mathbf{x}) = -4e^{2x_1+2x_2}$. For the diffusion matrix it holds $\alpha^{\text{ell}} = 1$ and $\alpha^{\text{cont}} < 8$ and the regularity assumptions are again fulfilled.
In this case we have the majorant

$$\begin{aligned} \mathcal{M}^2(v,\mathbf{y};\beta,\gamma) &:= (1+\beta)\,\|\mathbf{y}-\mathbf{A}\nabla v\|^2_{\mathbf{A}^{-1}} + \frac{1+\beta}{\beta}\frac{C^2_{F\Omega}}{\alpha^{\text{ell}}}\|f+\operatorname{div}\mathbf{y}\|^2_{L^2(\Omega)} \\ &= (1+\beta)\mathcal{M}^2_{\text{D}} + \frac{1+\beta}{\beta}\frac{C^2_{F\Omega}}{\alpha^{\text{ell}}}\mathcal{M}^2_{\text{Eq}}, \end{aligned} \tag{6.4}$$

from Section 2.4, which gives us a guaranteed upper bound for the energy error

$$Err_E := \|\nabla(u-u_h)\|_{\mathbf{A}}.$$

For problem (6.2), we can replace $\mathbf{A}$ by the identity matrix. In our computations we estimate the constant from the Friedrichs inequality by

$$C_{F\Omega} \le \frac{1}{\pi\sqrt{2}},$$

see (A.10).
The practical computation of this majorant was explained in Subsection 5.2.5, which gives two different algorithms for minimizing the majorant and getting a sharp upper bound. We will measure the sharpness with the efficiency index

$$\text{Eff}_{\mathcal{M}} := \frac{\mathcal{M}}{Err_E}.$$

Further, there are some parameters which we have to choose. We will state results with different parameters and strategies in order to compare them:

a) First, we consider Algorithm 2 and hence we use a gradient recovery procedure for the initial guess $\mathbf{y}_0 = \mathbf{A}G_h(\nabla u_h) = \mathbf{A}S^mQ_h(\nabla u_h)$. We can choose the value m, where we will test $m = 0,1,2$, according to the results of Section 6.1. Further, we can choose the number of iterations $I_{\max} \ge 0$.

b) Second, we consider Algorithm 3 with $\beta = 1$ as an initial guess. The approximation $\mathbf{y}_0$ is then computed by solving the linear system, so this procedure does not depend on the choice of the gradient recovery technique. We can again choose the number of iterations $I_{\max} \ge 1$, which should be kept rather small.

c) Finally, we consider Algorithm 2 without the parameters β and γ, according to Remark 5.2.

The results for these tests are listed in the following for the Dirichlet problem (6.2):

a) The values differ only slightly in the case where $I_{\max} = 0$ and where we vary $m = 0, 1, 2$. For $I_{\max} = 1$ and $I_{\max} = 2$, they are almost equal when varying m. Therefore, we show in Table 6.3 the results for $m = 0$:

		$I_{\max} = 0$		$I_{\max} = 1$		$I_{\max} = 2$	
nt	Err_E	$\mathcal{M}$	$\text{Eff}_{\mathcal{M}}$	$\mathcal{M}$	$\text{Eff}_{\mathcal{M}}$	$\mathcal{M}$	$\text{Eff}_{\mathcal{M}}$
4	1.25e+0	1.74e+0	1.39	1.47e+0	1.18	1.46e+0	1.17
16	6.04e-1	9.93e-1	1.64	7.64e-1	1.27	7.62e-1	1.26
64	2.99e-1	5.60e-1	1.87	3.83e-1	1.28	3.82e-1	1.28
256	1.49e-1	3.26e-1	2.19	1.91e-1	1.28	1.91e-1	1.28
1024	7.44e-2	1.97e-1	2.64	9.56e-2	1.28	9.56e-2	1.28
4096	3.72e-2	1.22e-1	3.29	4.78e-2	1.28	4.78e-2	1.28
16384	1.86e-2	7.85e-2	4.22	2.39e-2	1.28	2.39e-2	1.28
Order		0.75		0.99		0.99	

Table 6.3: Test a) for $m = 0$.

We observe that we need at least one minimization step in order to have a good error indicator. Further, $I_{\max} = 2$ cannot achieve an improvement, hence one minimization step would be our choice in this case. The efficiency index converges to a value 1.28, which is close to 1 but not optimal. In Figure 6.4 this number shows as a gap between the majorant and the exact error, further we depicted the behaviour of the terms $\mathcal{M}_{\text{D}}$ and $\mathcal{M}_{\text{Eq}}$.

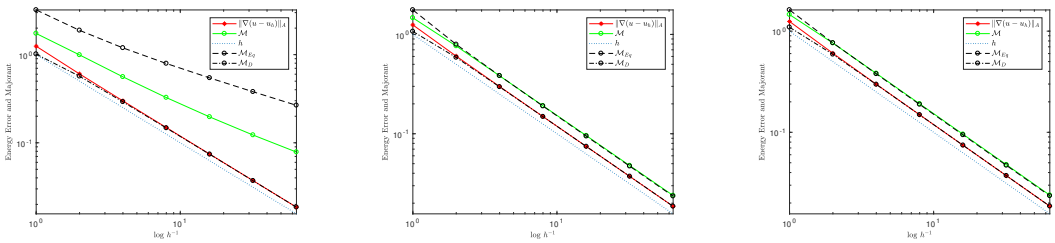


Figure 6.4: Test a) for $m = 0$, where $I_{\max} = 0$, $I_{\max} = 1$ and $I_{\max} = 2$ from left to right.

b) In this case $I_{\max}$ has to be at least one, to ensure that the flux approximation is in $H(\Omega, \text{div})$. One minimization step gives the efficiency index 1.47, which can be improved with one more step. For $I_{\max} = 2$, we get the efficiency index 1.28 and $I_{\max} = 3$ does not improve the convergence further. See also Table 6.4 and Figure 6.5.

		$I_{\max} = 1$		$I_{\max} = 2$		$I_{\max} = 3$	
nt	Err_E	$\mathcal{M}$	$\text{Eff}_{\mathcal{M}}$	$\mathcal{M}$	$\text{Eff}_{\mathcal{M}}$	$\mathcal{M}$	$\text{Eff}_{\mathcal{M}}$
4	1.25e+0	1.61e+0	1.29	1.46e+0	1.17	1.46e+0	1.17
16	6.04e-1	8.66e-1	1.43	7.62e-1	1.26	7.62e-1	1.26
64	2.99e-1	4.37e-1	1.46	3.82e-1	1.28	3.82e-1	1.28
256	1.49e-1	2.19e-1	1.47	1.91e-1	1.28	1.91e-1	1.28
1024	7.44e-2	1.09e-1	1.47	9.56e-2	1.28	9.56e-2	1.28
4096	3.72e-2	5.47e-2	1.47	4.78e-2	1.28	4.78e-2	1.28
16384	1.86e-2	2.74e-2	1.47	2.39e-2	1.28	2.39e-2	1.28
Order		0.99		0.99		0.99	

Table 6.4: Test b).

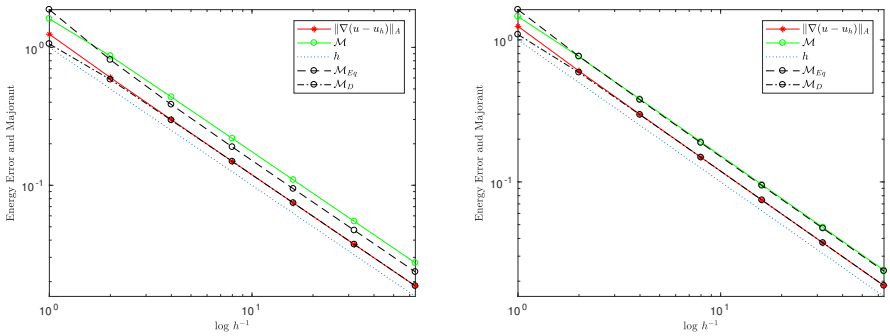


Figure 6.5: Test b), where $I_{\max} = 1$ (left) and $I_{\max} = 2$ (right).

c) With the same arguments as in Test a), we only list the results for $m = 0$, see Table 6.5 and Figure 6.6. We see that the efficiency index 1.47 is higher than 1.28 achieved in Test a), thus this strategy is not optimal, as expected.

		$I_{\max} = 0$		$I_{\max} = 1$		$I_{\max} = 2$	
nt	Err_E	$\mathcal{M}$	$\text{Eff}_{\mathcal{M}}$	$\mathcal{M}$	$\text{Eff}_{\mathcal{M}}$	$\mathcal{M}$	$\text{Eff}_{\mathcal{M}}$
4	1.25e+0	1.76e+0	1.41	1.61e+0	1.29	1.61e+0	1.29
16	6.04e-1	1.00e+0	1.66	8.66e-1	1.43	8.66e-1	1.43
64	2.99e-1	5.60e-1	1.88	4.37e-1	1.46	4.37e-1	1.46
256	1.49e-1	3.27e-1	2.20	2.19e-1	1.47	2.19e-1	1.47
1024	7.44e-2	2.02e-1	2.72	1.09e-1	1.47	1.09e-1	1.47
4096	3.72e-2	1.32e-1	3.54	5.47e-2	1.47	5.47e-2	1.47
16384	1.86e-2	8.87e-2	4.77	2.74e-2	1.47	2.74e-2	1.47
Order		0.72		0.99		0.99	

Table 6.5: Test c) for $m = 0$.

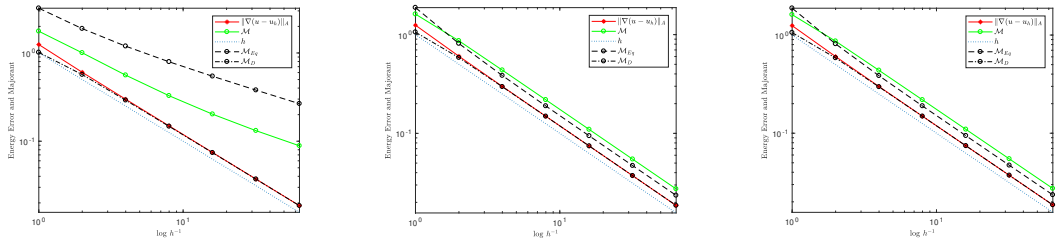


Figure 6.6: Test c) for $m = 0$, where $I_{\max} = 0$, $I_{\max} = 1$ and $I_{\max} = 2$ from left to right.

In conclusion, for the Dirichlet problem (6.2), we get a good error estimate if we choose Test a) with $I_{\max} = 1$ and $m = 0$, or Test b) with $I_{\max} = 2$. We did not get an asymptotically exact estimate, indicated by the efficiency index 1.28. This is not surprising, since we already mentioned in the beginning of Subsection 5.2.5, that we need some extra effort to get a good flux reconstruction. Meaning, that we get a good error estimate with employing one or two minimization steps, which results in a rather inexpensive procedure. To achieve an efficiency index 1 or close to 1, one would have to consider a dual mixed finite element method, where the reconstruction of the flux comes directly from the approximation.

An other possible explanation is, that the right-hand side function f is a smooth function, whereas $\operatorname{div} \mathbf{y}$ is only piecewise linear. Possible attempts to improve the estimate could be a different gradient recovery procedure which takes into account the equilibrium term $\mathcal{M}_{\text{Eq}}$, or to add a data oscillation term as explained in [9].

To illustrate this argument, consider the Dirichlet problem (6.3), where the right-hand side function f is similar to the one in (6.2), but with much higher values. This results in an efficiency index 3.40, i.e., the gap between the majorant and the energy error gets bigger and our estimate is less sharp. In Figure 6.7, we see that $\mathcal{M}_{\mathrm{Eq}}$ converges linearly and it does not hold $\mathcal{M}_{\mathrm{Eq}} \approx 0$, further we see the gap between the majorant and the exact error, while the asymptotic rate is preserved.

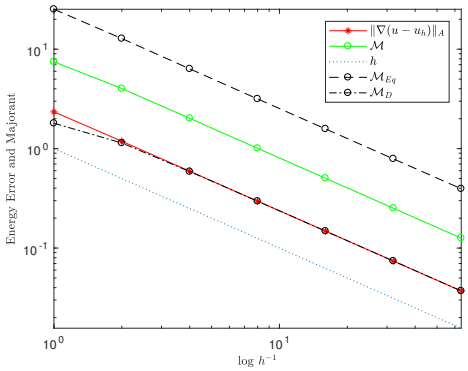


Figure 6.7: Test a) for problem (6.3), where $I_{\max} = 1$ and $m = 0$.

In the case of a constant right-hand side function f, it turns out that the efficiency index converges to 1. For that, consider the Dirichlet problem

$$\begin{aligned} -\operatorname{div}(\nabla u) &= f \quad \text{in } \Omega = (0,1)^2, \\ u &= g \quad \text{on } \Gamma, \end{aligned} \tag{6.5}$$

with a solution $u(\mathbf{x}) = x_1^2 - x_2^2$ depicted in Figure 6.8, hence $f(\mathbf{x}) \equiv 0$ and $g(\mathbf{x}) = x_1^2 - x_2^2$. In this case the regularity assumptions are not satisfied, but it fulfils the condition of having a constant right-hand side.

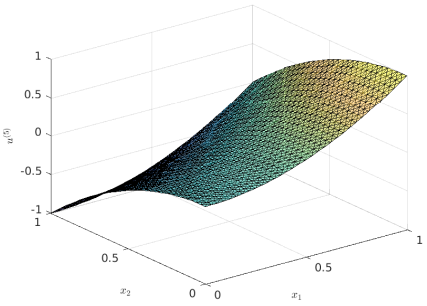


Figure 6.8: Approximated solution of problem (6.5).

		$I_{\max} = 0$		$I_{\max} = 1$		$I_{\max} = 2$	
nt	Err_E	$\mathcal{M}$	$\mathrm{Eff}_{\mathcal{M}}$	$\mathcal{M}$	$\mathrm{Eff}_{\mathcal{M}}$	$\mathcal{M}$	$\mathrm{Eff}_{\mathcal{M}}$
16	4.08e-1	7.39e-1	1.81	4.22e-1	1.03	3.62e-1	0.89
64	2.04e-1	5.33e-1	2.61	2.28e-1	1.12	1.99e-1	0.97
256	1.02e-1	3.59e-1	3.52	1.13e-1	1.11	1.01e-1	0.99
1024	5.10e-2	2.40e-1	4.71	5.52e-2	1.08	5.10e-2	1.00
4096	2.55e-2	1.62e-1	6.33	2.70e-2	1.06	2.55e-2	1.00
16384	1.28e-2	1.10e-1	8.59	1.32e-2	1.04	1.28e-2	1.00
Order		0.56		1.01		0.97	

Table 6.6: Test a) for $m = 0$ for problem (6.5).

Indeed, for problem (6.5) with constant right-hand side, we can achieve an efficiency index 1.00, for $I_{\max} = 2$ and $m = 0$, as shown in Table 6.6. Further, we observe the different behaviour of $\mathcal{M}_{\mathrm{Eq}}$ and the sharpness of the majorant in Figure 6.9.

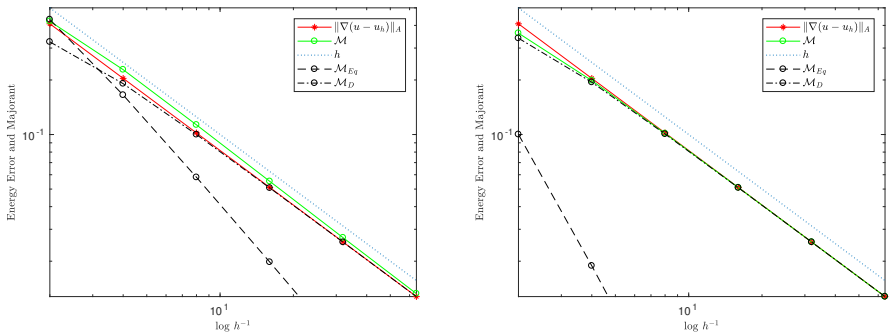


Figure 6.9: Test a), where $I_{\max} = 1$ (left) and $I_{\max} = 2$ (right) and $m = 0$ for problem (6.5).

6.3 Total Error Majorant

6.3.1 Quasi 1d Problem

In order to compare the behaviour of the majorants to the exact errors, we consider a quasi-one-dimensional example from [1], where the exact solutions can be computed.
Consider the homogeneous mixed Dirichlet-Neumann problem

$$\begin{aligned} -\operatorname{div}(\mathbf{A}_\varepsilon \nabla u_\varepsilon) &= f \quad \text{in } \Pi_{\mathbf{i}}^\varepsilon, \quad \forall \mathbf{i}, \\ u_\varepsilon &= 0 \quad \text{on } \Gamma_D := \{x_1 = 0\} \cup \{x_1 = 1\}, \\ \langle \mathbf{n}, \mathbf{A}_\varepsilon \nabla u_\varepsilon \rangle &= 0 \quad \text{on } \Gamma_N := \partial\Pi_{\mathbf{i}}^\varepsilon \backslash \Gamma_D, \end{aligned} \tag{6.6}$$

with the diffusion coefficient

$$\mathbf{A}_\varepsilon(\mathbf{x}) := \begin{pmatrix} 2 + \cos\left(2\pi \frac{x_1}{\varepsilon}\right) & 0 \\ 0 & 2 + \cos\left(2\pi \frac{x_1}{\varepsilon}\right) \end{pmatrix} \tag{6.7}$$

and $f(\mathbf{x}) \equiv 1$. We consider $\Omega = (0,1)^2$ and $\widehat{\Pi} = (0,1)^2$.

1) Then, we solve the following cell problem:

$$\begin{aligned} \operatorname{div}\left(\widehat{\mathbf{A}} \nabla \widehat{N}_j\right) &= \operatorname{div}\left(\widehat{\mathbf{A}}_j\right) \qquad \text{for } \mathbf{y} \in \widehat{\Pi}, \\ \widehat{N}_j \quad \widehat{\Pi} &- \text{periodic}, \\ \left\langle \widehat{N}_j \right\rangle_{\widehat{\Pi}} &= 0, \end{aligned}$$

for $j = 1, 2$ and

$$\widehat{\mathbf{A}}(\mathbf{y}) := \begin{pmatrix} 2 + \cos(2\pi y_1) & 0 \\ 0 & 2 + \cos(2\pi y_1) \end{pmatrix}.$$

The solutions of the cell problem are:

$$\widehat{N}_1(y_1) = y_1 - \sqrt{3} \int_0^{y_1} \frac{1}{2 + \cos(2\pi t)} \, \mathrm{d}t, \quad \widehat{N}_2(y_1) = 0.$$

The ellipticity and continuity constant are equal to $\widehat{\alpha}^{\mathrm{ell}} = 1$ and $\widehat{\alpha}^{\mathrm{cont}} = 3$. For $\alpha_\varepsilon^{\mathrm{ell}}$ we use the value of $\widehat{\alpha}^{\mathrm{ell}}$, which is a good estimate for rather large values $0 < \varepsilon < 1$. For $\varepsilon \to 0$, it holds $\alpha_\varepsilon^{\mathrm{ell}} \to 3$.

2) The homogenized coefficient is

$$\mathbf{A}_0 = \begin{pmatrix} \sqrt{3} & 0 \\ 0 & 2 \end{pmatrix}$$

and $\alpha_0^{\text{ell}} = \sqrt{3}$, $\alpha_0^{\text{cont}} = 2$.

3) The homogenized boundary value problem states

$$-\operatorname{div}(\mathbf{A}_0 \nabla u_0) = 1 \quad \text{in } \Omega.$$

Since we consider a quasi-one-dimensional problem, we solve the equivalent boundary value problem

$$-\partial_{x_1}\left((\mathbf{A}_0)_{1,1}\,\partial_{x_1} u_0\right) = 1 \quad \text{in } (0,1).$$

With the coefficient from above we have

$$-\sqrt{3}\partial_{x_1}(\partial_{x_1} u_0) = 1 \quad \text{in } (0,1),$$

which has the exact solution and the exact gradient

$$u_0(x_1) = -\frac{x_1^2}{2\sqrt{3}} + \frac{x_1}{2\sqrt{3}}, \quad \partial_{x_1} u_0(x_1) = -\frac{x_1}{\sqrt{3}} + \frac{1}{2\sqrt{3}}.$$

The boundary conditions are the following:

$$u_0(0, x_2) = u_0(1, x_2) = 0$$

for $x_2 \in [0,1]$ and

$$\left\langle \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \mathbf{A}_0 \begin{pmatrix} -\frac{x_1}{\sqrt{3}} + \frac{1}{2\sqrt{3}} \\ 0 \end{pmatrix} \right\rangle = 0, \quad \left\langle \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \mathbf{A}_0 \begin{pmatrix} -\frac{x_1}{\sqrt{3}} + \frac{1}{2\sqrt{3}} \\ 0 \end{pmatrix} \right\rangle = 0$$

for $x_1 \in [0,1]$.

4) The exact solution of the original problem (6.6) states

$$u_\varepsilon(x_1) = \int_0^{x_1} \frac{-t}{2 + \cos\left(2\pi \frac{t}{\varepsilon}\right)} \,\mathrm{d}t + \frac{1}{2} \int_0^{x_1} \frac{1}{2 + \cos\left(2\pi \frac{t}{\varepsilon}\right)} \,\mathrm{d}t$$

and the gradient is

$$\partial_{x_1} u_\varepsilon(x_1) = \frac{1}{2\left(2 + \cos\left(2\pi \frac{x_1}{\varepsilon}\right)\right)} - \frac{x_1}{2 + \cos\left(2\pi \frac{x_1}{\varepsilon}\right)}.$$

Since we consider a homogenized problem with mixed boundary conditions, we have to consider the Poincaré constant for both problems. From (A.10) we have $C_{P\Omega} = C_{P\widehat{\Omega}} \leq \frac{\sqrt{2}}{\pi}$. Moreover, the discretization majorant for u_0 consists of a third term $\|g - \langle \mathbf{y}, \mathbf{n} \rangle\|_{L^2(\Gamma)}$, due to the Neumann boundary condition. Since $g = 0$ in this case, we observe that this term is negligible.
The approximated solutions of $\widehat{N}_1$ and $\widehat{N}_2$ of the cell problem are depicted in Figure 6.10. We omit in the following results for $\widehat{N}_2$. The behaviour of the discretization majorant from Proposition 4.5 in comparison to the exact energy error is shown in Figure 6.11, where we used the gradient recovery with $m = 0$ and for the minimization procedure $I_{\max} = 1$. The efficiency index converges to 3.63, which is also visible in the gap between the majorant and the exact error. Hence, the majorant from Proposition 4.5 is indeed a guaranteed upper bound for the cell problem.

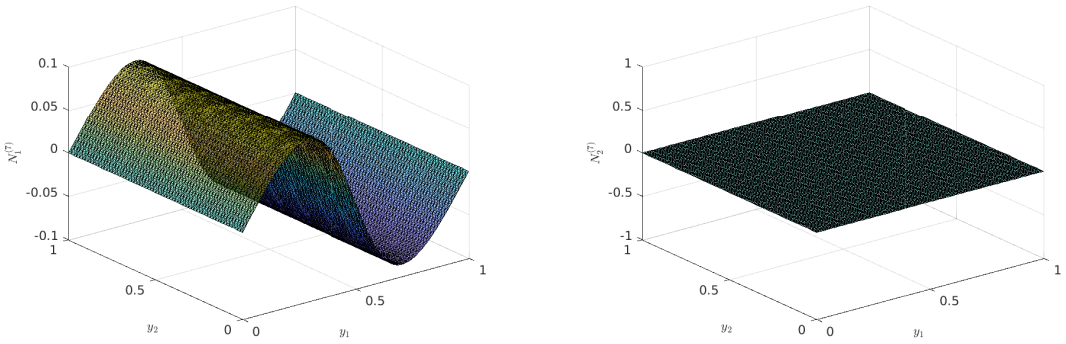


Figure 6.10: Approximated solution of $\widehat{N}_1$ and $\widehat{N}_2$.

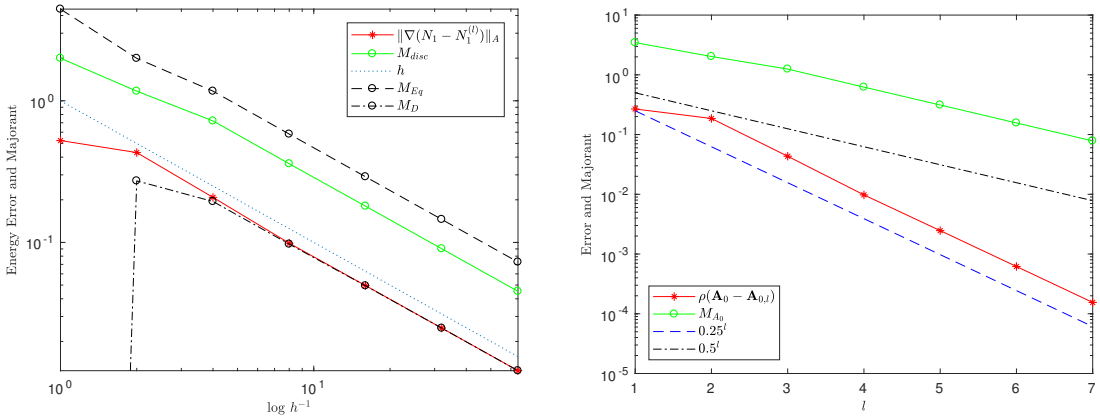


Figure 6.11: $\mathcal{M}_{\text{disc}}\left(\widehat{N}_1^{(l)}\right)$ for $m = 0$ and $I_{\max} = 1$ (left) and the approximation error and majorant for $\mathbf{A}_0$ (right).

By using the cell problems, we compute the approximated homogenized coefficient and its upper bound according to Proposition 4.6. From Figure 6.11, we observe that the asymptotic rate of the exact approximation error is faster than expected, still the majorant gives a guaranteed upper bound.
The approximated solution $u_0^{(l,j)}$ and the sharpness of the discretization majorant from Proposition 4.7 are depicted in Figure 6.12. For the same parameters as before, we achieve an efficiency index 1.04.

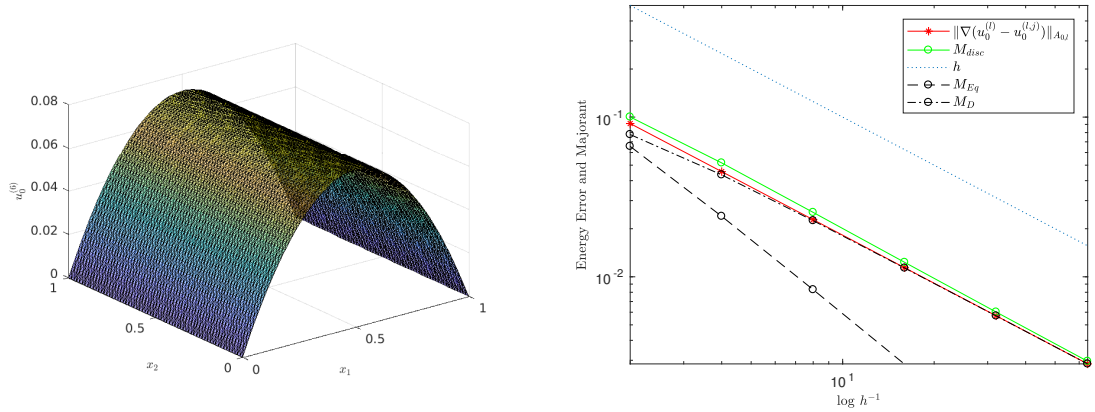
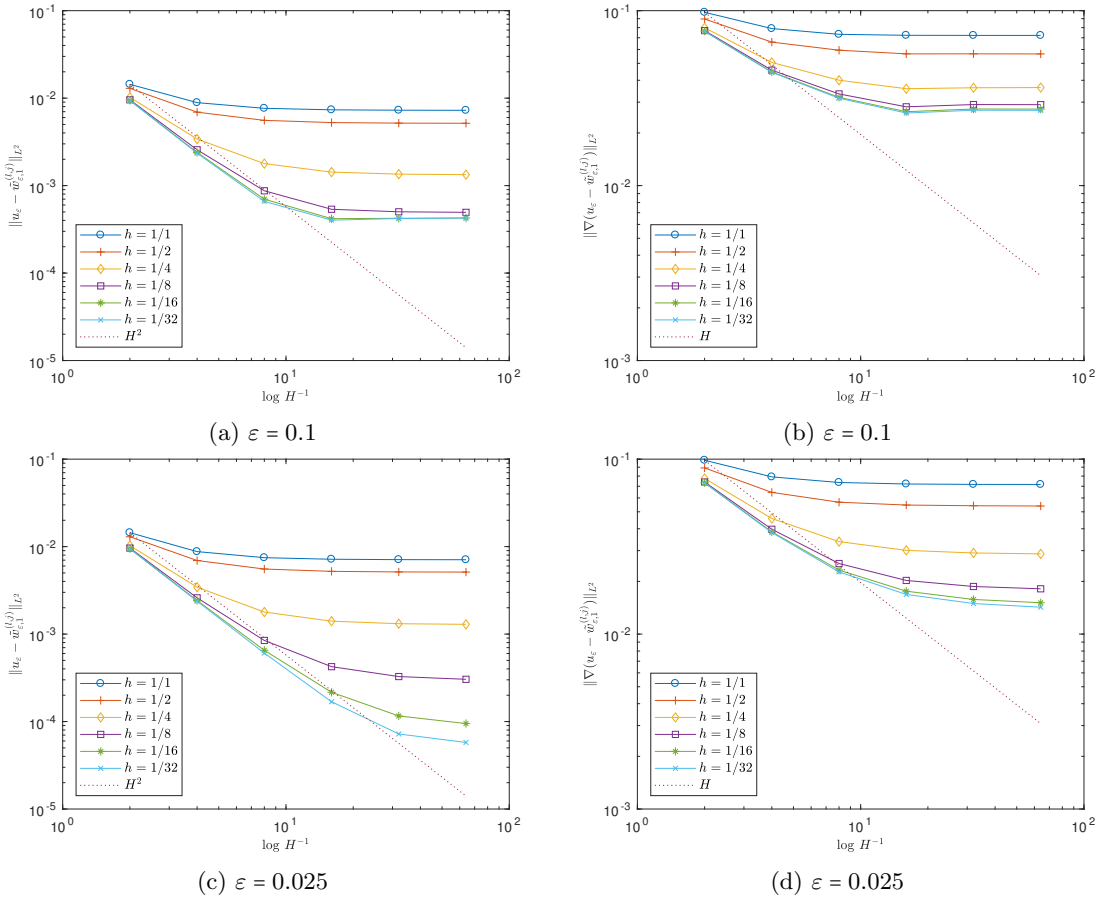


Figure 6.12: Approximated solution of $u_0^{(l)}$ and $\mathcal{M}_{\text{disc}}\left(u_0^{(l,j)}\right)$ for $m = 0$ and $I_{\max} = 1$.

For the two scale approximation, we examine in the following the convergence of the L^2-, H^1-, energy (EN) and flux (FL) error denoted by

$$L2 = \left\| u_\varepsilon - \widetilde{w}_{\varepsilon,1}^{(l,j)} \right\|_{L^2(\Omega)}, \qquad H1 = \left\| \nabla \left(u_\varepsilon - \widetilde{w}_{\varepsilon,1}^{(l,j)} \right) \right\|_{L^2(\Omega)},$$
$$EN = \left\| \nabla \left(u_\varepsilon - \widetilde{w}_{\varepsilon,1}^{(l,j)} \right) \right\|_{\mathbf{A}_\varepsilon}, \qquad FL = \left\| \mathbf{A}_{0,l} \nabla u_0^{(l,j)} - \mathbf{A}_\varepsilon \nabla \widetilde{w}_{\varepsilon,1}^{(l,j)} \right\|_{L^2(\Omega)}.$$

Recall the notation of the mesh $\mathcal{T}_{h_\varepsilon}$ for $\widetilde{w}_{1,\varepsilon}^{(l,j)}$, $\widehat{\mathcal{T}}_h$ for $\widehat{\mathbf{N}}^{(l)}$ and $\mathcal{T}_H$ for $u_0^{(l,j)}$, from Section 5.3.
In Figure 6.13, we see that the $L2$ error converges for a given h quadratically, i.e. with $O(H^2)$, until it reaches the point where the two scale approximation error dominates. This can be seen in the difference between the plots for $\varepsilon = 0.1$ and $\varepsilon = 0.025$. The same behaviour can be observed for the $H1$ error, with the difference that it converges linearly for a given h.



(a) $\varepsilon = 0.1$ (b) $\varepsilon = 0.1$

(c) $\varepsilon = 0.025$ (d) $\varepsilon = 0.025$

Figure 6.13: $L2$ and $H1$ error of the two scale approximation for different values of h with respect to H.

In order to observe the a priori convergence rate from Theorem 3.1, we compute the error terms with respect to ε and for a certain accuracy of the cell and homogenized problems. In the following test, $\widehat{\mathbf{N}}^{(l)}$ is computed on a mesh with 4096 elements, which corresponds to $h = 1/32$. The $H1$ and $L2$ errors are then 2.03e-2 and 1.70e-4 for $\widehat{N}_1$. For the computation of $\mathbf{A}_{0,l}$, we use the approximation of $\widehat{\mathbf{N}}^{(l)}$ on a mesh with 16384 elements and $h = 1/64$, then the exact approximation error is 1.55e-4. Further, $u_0^{(l,j)}$ is computed on a mesh with 16384 elements, which corresponds to $H = 1/64$. The $H1$ and $L2$ error are 2.13e-3 and 9.26e-6.
In Figure 6.14, we see the $O\left(\varepsilon^{1/2}\right)$-rate of the $H1$ error, as shown in Theorem 3.1. For the $L2$ error we would expect a $O(\varepsilon)$-rate, which is improved due to the Clément operator. Further, in Figure 6.15, we observe the $O\left(\varepsilon^{1/2}\right)$-rate of the flux term FL, which was also shown in Theorem 3.1.

In Figure 6.15, the approximated two scale approximation is depicted for the rather large value $\varepsilon = 0.2$. In this case, one can clearly see the oscillating behaviour in comparison to u_0. Since this example oscillates only in one direction, this can further be observed from a different view point as in Figure 6.16, where we additionally plotted the exact solution u_ε.
In conclusion, Figure 6.17 shows the behaviour of the energy error and the total error majorant from Theorem 4.11 for different values of h_ε. The mesh size for the cell and homogenized problem where chosen as $h = \frac{h_\varepsilon}{\varepsilon}$ and $H = h$. The energy error converges, as h_ε gets smaller, as expected. The total error majorant also gets smaller as h_ε does, but the asymptotic rate $O(\varepsilon^{1/2})$ is not achieved. Thus, the majorant is not optimal, but it gives a guaranteed upper bound.

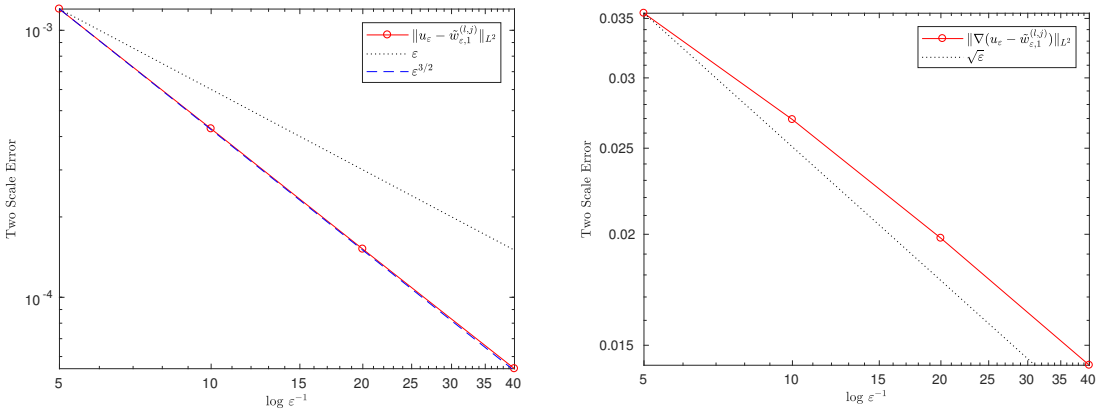


Figure 6.14: $L2$ and $H1$ error of the two scale approximation with respect to ε.

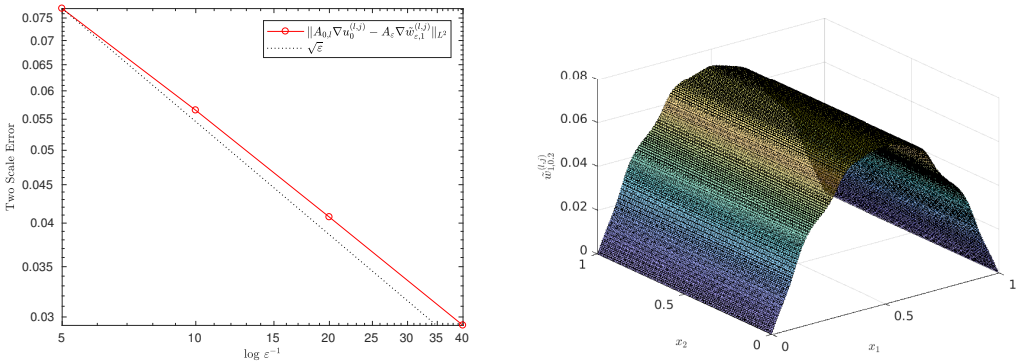
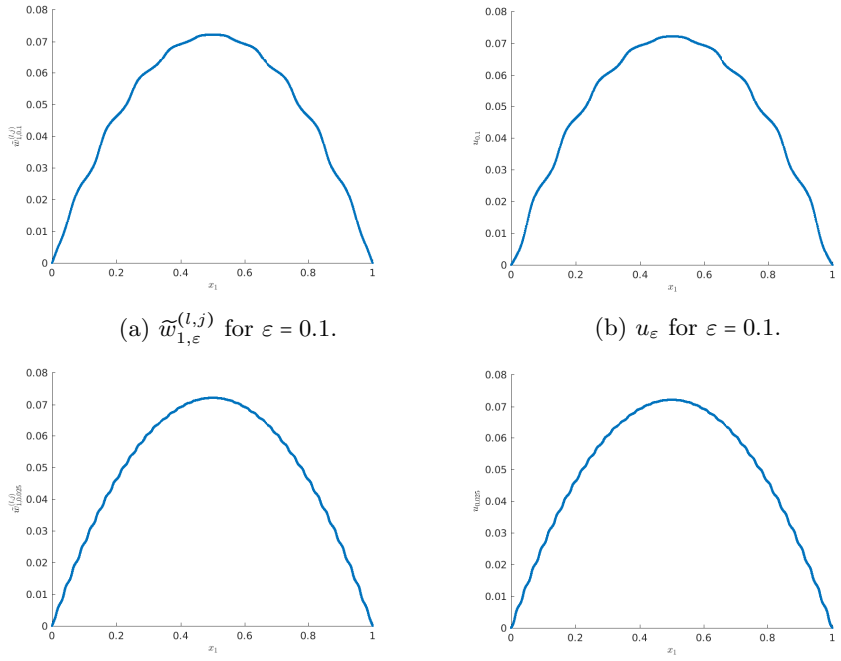


Figure 6.15: Computable error term FL of the two scale approximation with respect to ε (left) and approximated solution $\widetilde{w}_{1,\varepsilon}^{(l,j)}$ for $\varepsilon = 0.2$ (right).



(a) $\widetilde{w}_{1,\varepsilon}^{(l,j)}$ for $\varepsilon = 0.1$.

(b) u_ε for $\varepsilon = 0.1$.

(c) $\widetilde{w}_{1,\varepsilon}^{(l,j)}$ for $\varepsilon = 0.025$.

(d) u_ε for $\varepsilon = 0.025$.

Figure 6.16: Two scale approximation and exact solution for different values of ε.

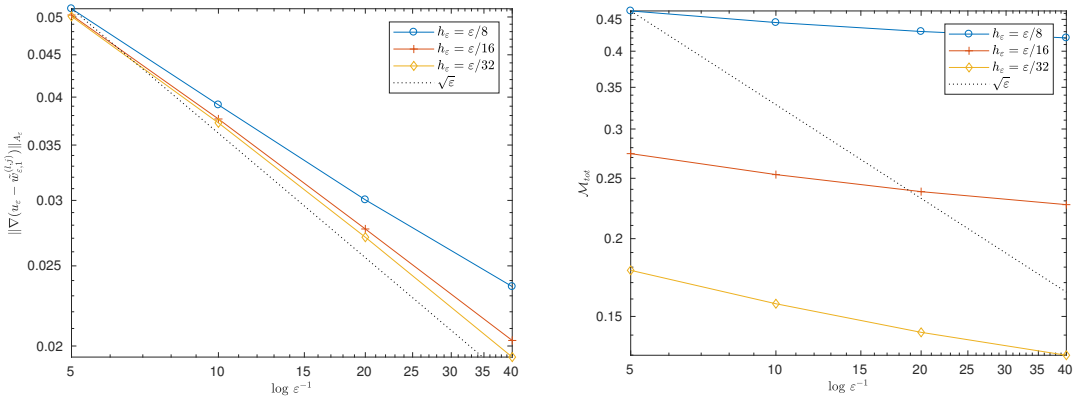


Figure 6.17: EN error and total error majorant of the two scale approximation for different values of h_ε with respect to ε.

6.3.2 2d Problem

Now we consider a two-dimensional example, where we cannot compute the exact solution:

$$\begin{aligned} -\operatorname{div}\left(\mathbf{A}_\varepsilon \nabla u_\varepsilon\right) + c_\varepsilon u_\varepsilon &= f \quad \text{in } \Pi_{\mathbf{i}}^\varepsilon, \quad \forall \mathbf{i}, \\ u_\varepsilon &= 0 \quad \text{on } \Gamma. \end{aligned} \tag{6.8}$$

As before, let $\Omega = (0,1)^2$ and $\widehat{\Pi} = (0,1)^2$. The coefficients and the function f are inspired by the two-dimensional problem from [31] and defined by:

$$\mathbf{A}_\varepsilon(\mathbf{x}) := \begin{pmatrix} 6 + \cos\left(2\pi\frac{x_1}{\varepsilon}\right) + \cos\left(2\pi\frac{x_2}{\varepsilon}\right) & 0 \\ 0 & 6 + \cos\left(2\pi\frac{x_1}{\varepsilon}\right) + \cos\left(2\pi\frac{x_2}{\varepsilon}\right) \end{pmatrix}, \tag{6.9}$$

$c_\varepsilon(\mathbf{x}) \equiv 1$ and $f(\mathbf{x}) = 10x_1 + 10x_2$.

1) Then, we solve the following cell problem:

$$\begin{aligned} \operatorname{div}\left(\widehat{\mathbf{A}} \nabla \widehat{N}_j\right) &= \operatorname{div}\left(\widehat{\mathbf{A}}_j\right) \qquad \text{for } \mathbf{y} \in \widehat{\Pi}, \\ \widehat{N}_j &\quad \widehat{\Pi}\text{ – periodic }, \\ \left\langle \widehat{N}_j \right\rangle_{\widehat{\Pi}} &= 0, \end{aligned}$$

for $j = 1, 2$ and

$$\widehat{\mathbf{A}}(\mathbf{y}) := \begin{pmatrix} 6 + \cos(2\pi y_1) + \cos(2\pi y_2) & 0 \\ 0 & 6 + \cos(2\pi y_1) + \cos(2\pi y_2) \end{pmatrix}.$$

The right-hand side functions are

$$\begin{aligned} f_1(\mathbf{y}) &:= -\operatorname{div}(\widehat{\mathbf{A}}_1) = 2\pi \sin(2\pi y_1), \\ f_2(\mathbf{y}) &:= -\operatorname{div}(\widehat{\mathbf{A}}_2) = 2\pi \sin(2\pi y_2). \end{aligned}$$

The ellipticity and continuity constant are equal to $\widehat{\alpha}^{\text{ell}} = 4$ and $\widehat{\alpha}^{\text{cont}} = 8$. The solutions of the cell problem are not known.

2) The exact homogenized matrix is not known, but we can compute $c_0 = 1$.

3) The homogenized boundary value problem states

$$\begin{aligned} -\operatorname{div}\left(\mathbf{A}_0 \nabla u_0\right) + c_0 u_0 &= f \quad \text{in } \Omega, \\ u_0 &= 0 \quad \text{on } \Gamma, \end{aligned}$$

where the exact solution is not known, since we do not know $\mathbf{A}_0$.

4) The exact solution of the original problem (6.8) is not known. The ellipticity and continuity constant are estimated by $\alpha_\varepsilon^{\text{ell}} \geq 4$ and $\alpha_\varepsilon^{\text{cont}} = 8$.

For this example, we consider $C_{F\Omega} \leq \frac{1}{\pi\sqrt{2}}$ and $C_{P\widehat{\Pi}}$ as before. The approximated solutions of $\widehat{N}_1$ and $\widehat{N}_2$ of the cell problem are depicted in Figure 6.18. We can see that $\widehat{N}_1$ is similar to the quasi 1d problem, but this time with a non-constant periodic behaviour in the y_2-direction. $\widehat{N}_2$ shows the same behaviour, but with interchanged directions.
The behaviour of the discretization majorant from Proposition 4.5 is shown in Figure 6.19. In Figure 6.20, the estimate for the approximation error from Proposition 4.6 is depicted. The computed approximated homogenized coefficients are

$$\mathbf{A}_{0,l} = \begin{pmatrix} 5.92 & 0.00 \\ 0.00 & 5.92 \end{pmatrix}$$

and $c_0 = 1.00$.

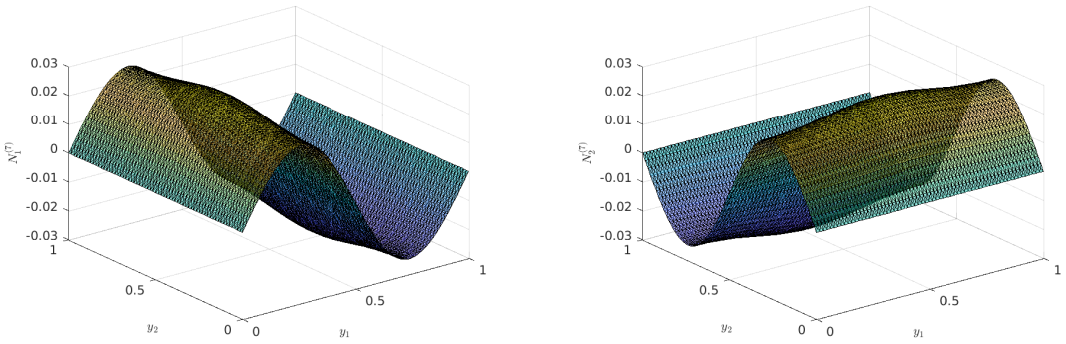


Figure 6.18: Approximated solution of $\widehat{N}_1$ and $\widehat{N}_2$.

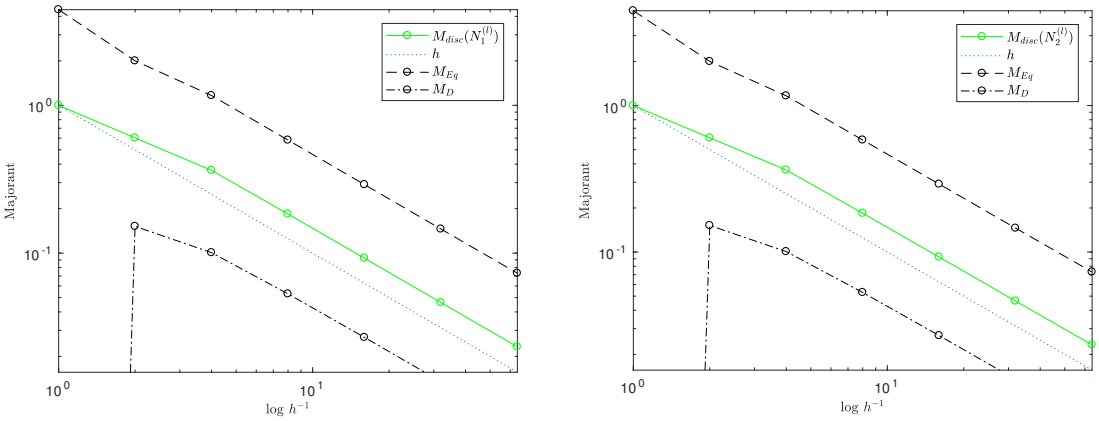


Figure 6.19: $\mathcal{M}_{\text{disc}}\left(\widehat{N}_1^{(l)}\right)$ and $\mathcal{M}_{\text{disc}}\left(\widehat{N}_2^{(l)}\right)$ for $m = 0$ and $I_{\max} = 1$.

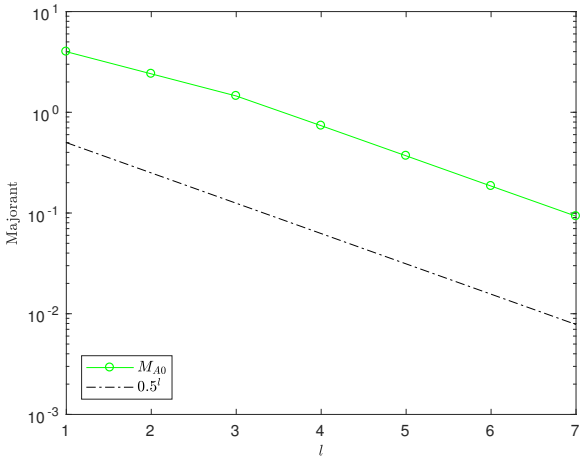


Figure 6.20: Approximation majorant for $\mathbf{A}_0$.

The approximated solution of $u_0^{(l)}$ and the behaviour of the discretization majorant from Proposition 4.7 are presented in Figure 6.21.
For the two scale approximation, we can no longer compare the majorant to the exact energy error, since we do not have the exact solution. Still, we can examine the performance of the flux error (FL) and the total error majorant. Figure 6.22 shows the behaviour of the total error majorant from Theorem 4.18 again for different values of h_ε, where $B_0 = 0$. As in the quasi 1d example, the asymptotic rate $O(\varepsilon^{1/2})$ is not achieved, but the majorant gets smaller as h_ε does.

Further, we see in Figure 6.22, that the a priori rate for the flux term FL is not preserved. This is not surprising, since the homogenized flux converges only weakly to the two scale flux (see [13]), i.e., the a priori rate is only realistic in theory or for simple examples. The flux term contributes to the total error majorant, which explains the slightly slower convergence rate of the majorant compared to the quasi 1d problem. Moreover, as depicted in Figure 6.23, the total error majorant does not capture the linear convergence for a fixed value of h and with respect to H.

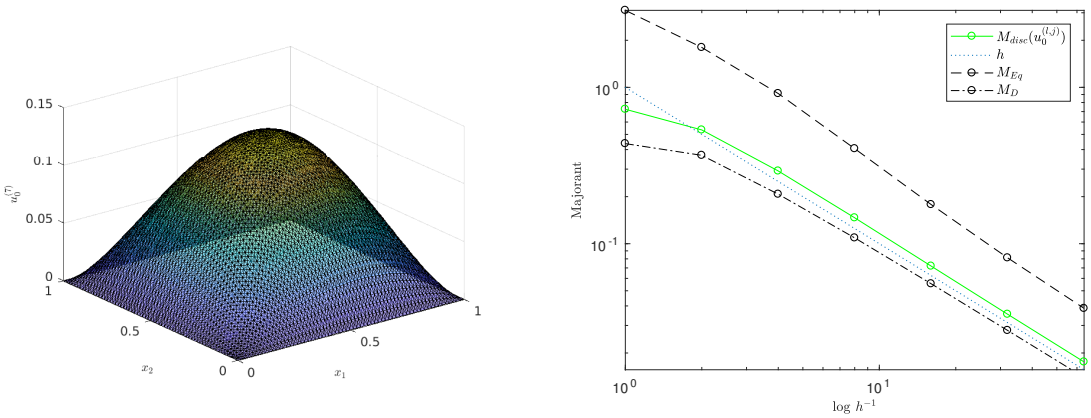


Figure 6.21: Approximated solution of $u_0^{(l)}$ and $\mathcal{M}_{\text{disc}}\left(u_0^{(l,j)}\right)$ for $m = 0$ and $I_{\max} = 1$.

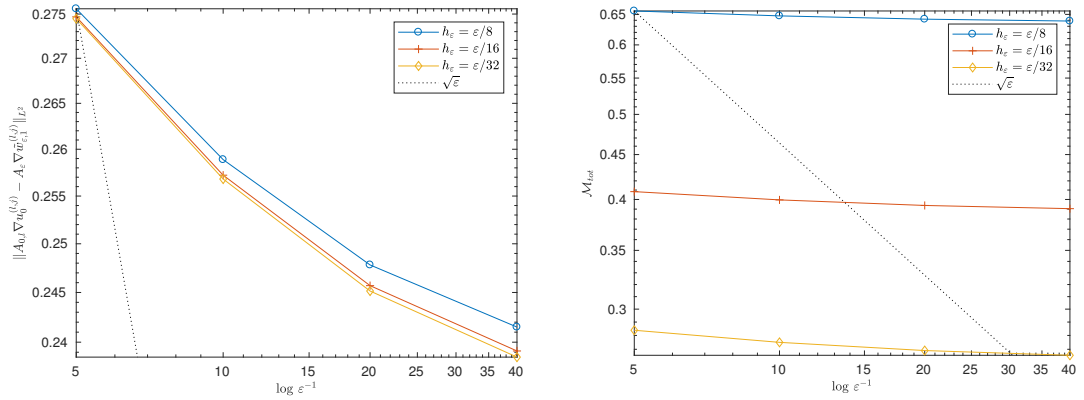
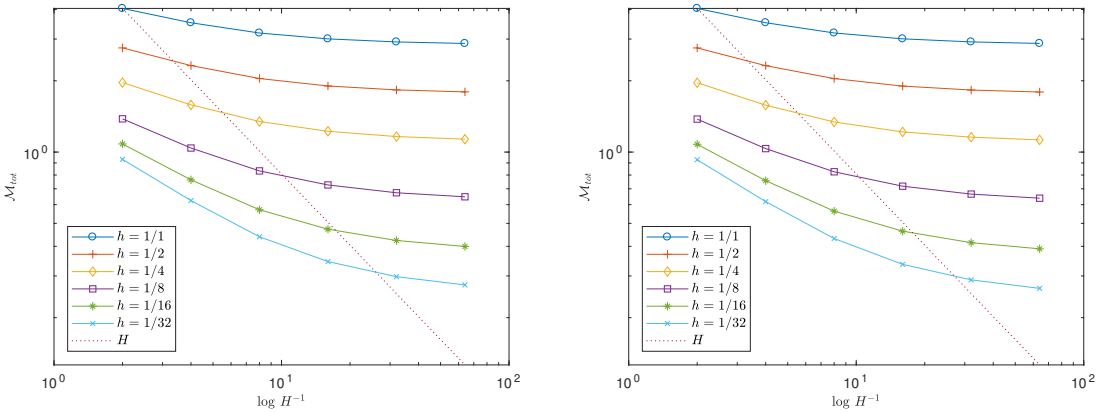


Figure 6.22: Computable error term FL and total error majorant of the two scale approximation for different values of h_ε, with respect to ε.



(a) $\varepsilon = 0.1$ (b) $\varepsilon = 0.025$

Figure 6.23: Total error majorant for different values of h with respect to H.

6.3.3 Oscillatory 2d Problem

Now we consider a two-dimensional example with a rapidly oscillating coefficient, where we cannot compute the exact solution:

$$\begin{aligned} -\operatorname{div}\left(\mathbf{A}_\varepsilon \nabla u_\varepsilon\right) + c_\varepsilon u_\varepsilon &= f \quad \text{in } \Pi_{\mathbf{i}}^\varepsilon, \quad \forall \mathbf{i}, \\ u_\varepsilon &= 0 \quad \text{on } \Gamma. \end{aligned} \tag{6.10}$$

As before, let $\Omega = (0,1)^2$ and $\widehat{\Pi} = (0,1)^2$. The function f and the coefficient c_ε are the same as in the 2d problem. The diffusion coefficient is chosen such that it is rapidly oscillatory and uniformly elliptic:

$$\mathbf{A}_\varepsilon(\mathbf{x}) := \begin{pmatrix} 1+\mu\left(\cos\left(2\pi\frac{x_1}{\varepsilon}\right)^2 + \cos\left(2\pi\frac{x_2}{\varepsilon}\right)^2\right) & 0 \\ 0 & 1+\mu\left(\cos\left(2\pi\frac{x_1}{\varepsilon}\right)^2 + \cos\left(2\pi\frac{x_2}{\varepsilon}\right)^2\right) \end{pmatrix}, \tag{6.11}$$

$\mu = 50$, $c_\varepsilon(\mathbf{x}) \equiv 1$ and $f(\mathbf{x}) = 10x_1 + 10x_2$.

1) Then, we solve the following cell problem:

$$\begin{aligned} \operatorname{div}\left(\widehat{\mathbf{A}}\nabla\widehat{N}_j\right) &= \operatorname{div}\left(\widehat{\mathbf{A}}_j\right) \qquad \text{for } \mathbf{y} \in \widehat{\Pi}, \\ \widehat{N}_j \quad & \widehat{\Pi}-\text{periodic}, \\ \left\langle \widehat{N}_j \right\rangle_{\widehat{\Pi}} &= 0, \end{aligned}$$

for $j = 1, 2$ and

$$\widehat{\mathbf{A}}(\mathbf{y}) := \begin{pmatrix} 1+\mu\left(\cos(2\pi y_1)^2 + \cos(2\pi y_2)^2\right) & 0 \\ 0 & 1+\mu\left(\cos(2\pi y_1)^2 + \cos(2\pi y_2)^2\right) \end{pmatrix}.$$

The right-hand side functions are

$$\begin{aligned} f_1(\mathbf{y}) &= 4\pi\mu\cos(2\pi y_1)\sin(2\pi y_1), \\ f_2(\mathbf{y}) &= 4\pi\mu\cos(2\pi y_2)\sin(2\pi y_2). \end{aligned}$$

The ellipticity and continuity constant are equal to $\widehat{\alpha}^{\text{ell}} = 1$ and $\widehat{\alpha}^{\text{cont}} = 2\mu + 1$. The solutions of the cell problem are not known.

2) The exact homogenized matrix is not known, but we can compute $c_0 = 1$.

3) The homogenized boundary value problem states

$$\begin{aligned} -\operatorname{div}\left(\mathbf{A}_0 \nabla u_0\right) + c_0 u_0 &= f \quad \text{in } \Omega, \\ u_0 &= 0 \quad \text{on } \Gamma, \end{aligned}$$

where the exact solution is not known, since we do not know $\mathbf{A}_0$.

4) The exact solution of the original problem (6.10) is not known. The ellipticity and continuity constant are estimated by $\alpha_\varepsilon^{\text{ell}} \geq 1$ and $\alpha_\varepsilon^{\text{cont}} = 2\mu + 1$.

We consider $C_{F\Omega}$ and $C_{P\widehat{\Pi}}$ as in the 2d problem. The approximated solutions of $\widehat{N}_1$ and $\widehat{N}_2$ of the cell problem are depicted in Figure 6.24, where we can see the effect of the constant $\mu = 50$.
The behaviour of the majorants is similar to the 2d problem, with the difference that the majorants now contain rather large constants due to the factor μ, therefore we leave them away.
The computed approximated homogenized coefficients are

$$\mathbf{A}_{0,l} = \begin{pmatrix} 43.62 & -0.00 \\ -0.00 & 43.62 \end{pmatrix} \quad \text{and} \quad c_0 = 1.00.$$

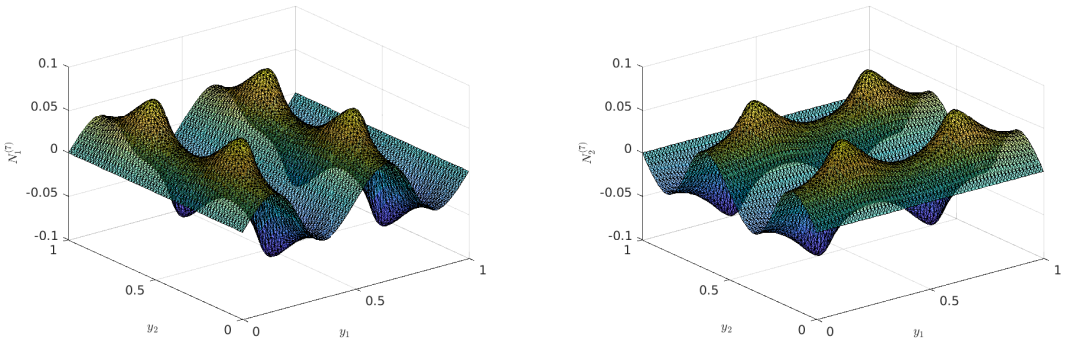


Figure 6.24: Approximated solution of $\widehat{N}_1$ and $\widehat{N}_2$.

The approximated solution of $u_0^{(l)}$ is depicted in Figure 6.25.

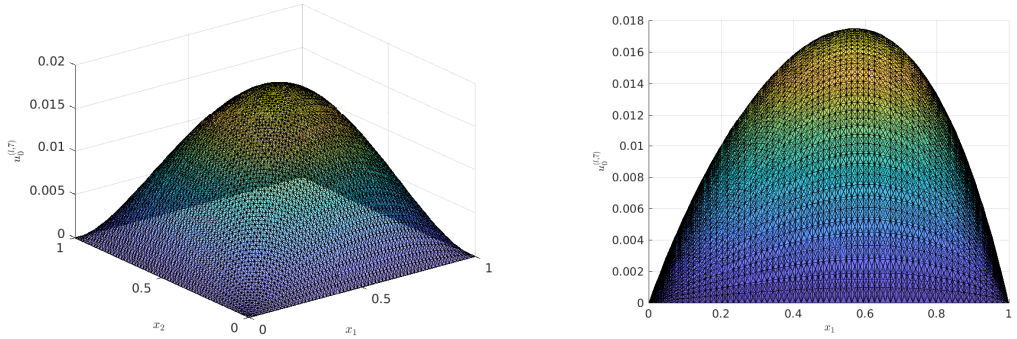


Figure 6.25: Approximated solution of $u_0^{(l)}$.

In Figure 6.26, the approximated two scale approximation is depicted for the rather large value $\varepsilon = 0.2$. One can clearly see the oscillating behaviour due to the choice of $\mathbf{A}_\varepsilon$. This can further be observed in comparison to u_0 and from a different view point.

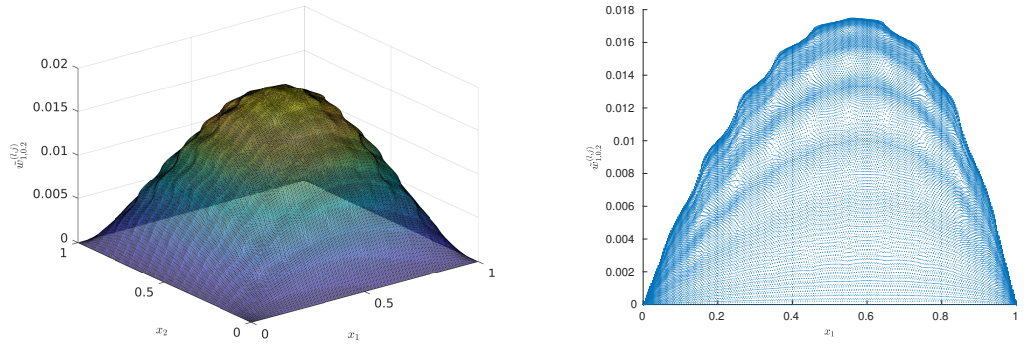


Figure 6.26: Approximated solution $\widetilde{w}_{1,\varepsilon}^{(l,j)}$ for $\varepsilon = 0.2$.

7 Conclusion

In this thesis, we presented an a posteriori error estimate for elliptic homogenization problems in the context of periodic structures. For that, we explained the cell problem and the homogenized problem, whose combination results in the two scale approximation. Then, we studied the discretization majorant for the cell problem and the modelling/discretization error for the homogenized problem in detail. As the main result, we developed a total error majorant, consisting of the discretization majorants for both problems and an additional term related to homogenization. This error estimate is a fully computable and guaranteed upper bound for the energy error of the two scale approximation, which includes all appearing error terms. Moreover, we suggested an error estimation strategy to balance the error terms.
The minimization procedure for the majorant, including a suitable gradient recovery procedure, is explained step by step. Further, the process of implementing each problem is described.
In the numerical experiments, we illustrated the efficiency and superconvergence of the chosen gradient recovery procedure. Further, we observed the accuracy and sharpness of the majorants for the cell and homogenized problem. It turns out, that the approximation estimate for the homogenized coefficient is not sharp and could be improved, still it gives a guaranteed upper bound. Moreover, we observed that the a priori convergence rate of the flux variables for the two scale approximation cannot be seen in every example and that the convergence rate of the L^2-error of the two scale approximation is improved due to the smoothing with the Clément operator. The derived total error majorant indeed gives a guaranteed upper bound for the energy error.
Future research could be done on the usage of this majorant as an error indicator. Further, the estimate of the approximation error of the homogenized coefficient could be improved. Finally, a gradient recovery procedure more adapted to this problem could be investigated, or a dual mixed finite element method could be applied.

A Mathematical Background

In this chapter we will shortly summarize important inequalities and tools from functional analysis which where needed before. The proofs and further details can be found in, e.g., [26], [10], [3], [9], [21], [13] and [4].

A.1 Vectors and Matrices

For vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$ of dimension $d \in \mathbb{N}$, the scalar product is defined by

$$\langle \mathbf{v}, \mathbf{w} \rangle := \sum_{i=1}^{d} v_i w_i.$$

For matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$, the scalar product is defined as the Frobenius scalar product:

$$\mathbf{A} : \mathbf{B} := \sum_{i=1}^{d}\sum_{j=1}^{d} a_{i,j} b_{i,j}.$$

For $a, b \in \mathbb{R}_{>0}$ and some $\beta > 0$, for $p, q \in [1, \infty]$ such that $\frac{1}{p} + \frac{1}{q} = 1$, we have **Young's inequality**:

$$ab \le \frac{1}{p}(\beta a)^p + \frac{1}{q}\left(\frac{b}{\beta}\right)^q. \tag{A.1}$$

We will mostly use Young's inequality for $p = q = 2$, which also holds for a pair of vectors and matrices:

$$2\langle \mathbf{v}, \mathbf{w} \rangle \le \beta \|\mathbf{v}\|^2 + \frac{1}{\beta}\|\mathbf{w}\|^2, \quad 2\mathbf{A} : \mathbf{B} \le \beta \|\mathbf{A}\|^2 + \frac{1}{\beta}\|\mathbf{B}\|^2,$$

where the norms are defined by

$$\|\mathbf{v}\| := \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}, \quad \|\mathbf{A}\| := \sqrt{\mathbf{A} : \mathbf{A}}.$$

From these inequalities follow these two useful inequalities:

$$\|\mathbf{v} + \mathbf{w}\|^2 \le (1+\beta)\|\mathbf{v}\|^2 + \frac{1+\beta}{\beta}\|\mathbf{w}\|^2, \quad \|\mathbf{A} + \mathbf{B}\|^2 \le (1+\beta)\|\mathbf{A}\|^2 + \frac{1+\beta}{\beta}\|\mathbf{B}\|^2. \tag{A.2}$$

For a Hilbert space H with scalar product $(\cdot,\cdot)_H$ and associated norm $\|\cdot\|_H$, which are explained in Section A.2, one can easily extend these inequalities to elements of H.
In general, for a constant matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ and any vector norm $\|\cdot\|$ on $\mathbb{R}^d$, we define the induced matrix norm by

$$\|\mathbf{A}\| := \sup_{\mathbf{v}\in\mathbb{R}^d\setminus\{\mathbf{0}\}} \frac{\|\mathbf{A}\mathbf{v}\|}{\|\mathbf{v}\|}$$

and by $\rho(\mathbf{A})$ we denote the spectral radius of $\mathbf{A}$. For $p \in \mathbb{R}_{\ge 1}$ and any vector $\mathbf{v} \in \mathbb{R}^d$, $\|\mathbf{v}\|_p$ denotes the standard p-norm. Then, the induced matrix norm is defined by

$$\|\mathbf{A}\|_p := \sup_{\mathbf{v}\in\mathbb{R}^d\setminus\{\mathbf{0}\}} \frac{\|\mathbf{A}\mathbf{v}\|_p}{\|\mathbf{v}\|_p}.$$

For matrix functions $\mathbf{A} \in L^\infty(\Omega, \mathbb{R}^{d\times d})$, we set

$$\|\mathbf{A}\|_{p,\Omega} := \operatorname*{ess\,sup}_{\mathbf{x}\in\Omega} \|\mathbf{A}(\mathbf{x})\|_p$$

and

$$\rho_\Omega(\mathbf{A}) := \operatorname*{ess\,sup}_{\mathbf{x}\in\Omega} \rho(\mathbf{A}(\mathbf{x})).$$

In general, it holds $\|\mathbf{A}\|_2 \le \rho(\mathbf{A})$. For $\mathbf{A} \in \mathbb{R}^{d\times d}_{\text{sym}}$, it further holds $\|\mathbf{A}\|_2 = \rho(\mathbf{A})$. Both statements are valid for matrix functions.
Considering the Frobenius norm $\|\mathbf{A}\|_F := \sqrt{\sum_{i,j=1}^d |a_{i,j}|^2}$, we have the following inequality:

$$\|\mathbf{A}\|_2 \le \|\mathbf{A}\|_F, \tag{A.3}$$

which follows from $\|\mathbf{A}\mathbf{v}\|_2 \le \|\mathbf{A}\|_F \|\mathbf{v}\|_2$ and the definition of $\|\mathbf{A}\|_2$.
For a differentiable function $w : \mathbb{R}^d \to \mathbb{R}$, the **gradient** of w is defined as

$$\nabla w = (\partial_{x_1} w, \partial_{x_2} w, \dots, \partial_{x_d} w)^\top,$$

where we used the shorthand notation $\partial_{x_i} w$ for the partial derivatives $\frac{\partial}{\partial x_i} w(\mathbf{x})$. For a differentiable vector function $\mathbf{v} : \mathbb{R}^d \to \mathbb{R}^d$, the **divergence** of $\mathbf{v}$ is defined as

$$\operatorname{div} \mathbf{v} = \sum_{i=1}^d \partial_{x_i} v_i.$$

The gradient of $\mathbf{v}$ denotes the **Jacobian matrix**

$$\nabla \mathbf{v} = \left(\partial_{x_j} v_i\right)_{i,j=1}^d.$$

Note that it holds, for w two times differentiable:

$$\operatorname{div}(\nabla w) = \Delta w = \sum_{i=1}^d \partial_{x_i}^2 w.$$

The following identity holds for $u : \mathbb{R}^d \to \mathbb{R}$ and w as before:

$$\operatorname{div}(\nabla u w) = \operatorname{div}(\nabla u)\, w + \langle \nabla u, \nabla w \rangle. \tag{A.4}$$

In the following we state several important integral identities.

Theorem A.1 (Gaussian integral theorem). *For a closed surface $\Gamma = \partial\Omega$, it holds:*

$$\int_\Omega \operatorname{div}(\mathbf{v})\, \mathrm{d}\mathbf{x} = \int_\Gamma \langle \mathbf{v}, \mathbf{n} \rangle \, \mathrm{d}s.$$

Here, $\mathbf{n} : \mathbb{R}^d \to \mathbb{R}^d$ denotes the outer normal direction at $\mathbf{x} \in \Gamma$, which is a unit vector. Then, the value $\langle \nabla u, \mathbf{n} \rangle$ denotes the normal derivative of u, i.e. $\frac{\partial}{\partial n} u = \langle \nabla u, \mathbf{n} \rangle$.
With identity (A.4) and Theorem A.1, **Green's first identity** follows:

$$\int_\Omega \operatorname{div}(\nabla u w)\, \mathrm{d}\mathbf{x} = \int_\Omega \left(\Delta u w + \langle \nabla u, \nabla w \rangle\right) \mathrm{d}\mathbf{x} = \int_\Gamma \langle \nabla u, \mathbf{n} \rangle w \, \mathrm{d}s. \tag{A.5}$$

Note that the right-hand side is zero, if w vanishes on the boundary.

A.2 Sobolev Spaces

We consider an open and connected subset $\Omega \subset \mathbb{R}^d$, which we call **domain**. Further, we assume that Ω is bounded and has a **Lipschitz boundary**, which is the case, if there exists $N \in \mathbb{N}$ and open sets $U_1, \dots, U_N \subset \mathbb{R}^d$ satisfying:

a) $\partial\Omega \subset \bigcup_{i=1}^N U_i$,

b) $\partial\Omega \cap U_i$ can be represented as the graph of a Lipschitz function for every $i = 1, \dots, N$.

The Lebesgue measure of a subset $\omega \subset \Omega$ is denoted by $|\omega| = \int_\omega 1 \, \mathrm{d}\mathbf{x}$. Note that all function spaces considered in this thesis are real-valued.

Definition A.2 (L^p spaces). *Let $p \in \mathbb{R}$ with $1 \leq p < \infty$, we define*

$$L^p(\Omega) := \left\{ f : \Omega \to \mathbb{R} \mid f \text{ is measurable and } \int_\Omega |f|^p \, \mathrm{d}\mathbf{x} < \infty \right\},$$

with the norm

$$\|f\|_{L^p(\Omega)} := \left(\int_\Omega |f(\mathbf{x})|^p \, \mathrm{d}\mathbf{x} \right)^{1/p}.$$

For $p = \infty$, we define

$$L^\infty(\Omega) := \{ f : \Omega \to \mathbb{R} \mid f \text{ is measurable and there exists } C \text{ s.t. } |f(\mathbf{x})| \leq C \text{ a.e. on } \Omega \},$$

with the norm

$$\|f\|_{L^\infty(\Omega)} := \operatorname*{ess\,sup}_{\mathbf{x} \in \Omega} \{ |f(\mathbf{x})| \}.$$

Definition A.3 (Conjugate exponent). *Let $1 \leq p \leq \infty$, we denote by q the conjugate exponent of p, if it holds:*

$$\frac{1}{p} + \frac{1}{q} = 1.$$

The convention $\frac{1}{\infty} = 0$ is assumed.

Theorem A.4.

a) L^p is a Banach space for any p, $1 \leq p \leq \infty$.

b) L^2 is a Hilbert space with scalar product

$$(f, g)_{L^2(\Omega)} := \int_\Omega f(\mathbf{x}) g(\mathbf{x}) \, \mathrm{d}\mathbf{x} \quad \text{for } f, g \in L^2(\Omega)$$

and the norm is defined by $\|f\|^2_{L^2(\Omega)} = (f, f)_{L^2(\Omega)}$.

Definition A.5 (Dual space). *The dual space of any vector space X is defined as*

$$X' := \{ l : X \to \mathbb{R} \mid l \text{ being a bounded and continuous functional} \}.$$

The corresponding norm is defined by

$$\|l\|_{X'} = \sup_{x \in X \setminus \{0\}} \frac{l(x)}{\|x\|_X}.$$

Theorem A.6 (Riesz representation theorem). *Let X be a Hilbert space with scalar product $(\cdot, \cdot)_X$. Then, $l \in X'$ if and only if there exists $x \in X$ such that for every $y \in X$ we have*

$$l(y) = (y, x)_X$$

and in this case $\|l\|_{X'} = \|x\|_X$. Moreover, x is uniquely determined by $l \in X'$.

Proposition A.7 (Dual space of L^p). *Let $1 \leq p < \infty$ and q its conjugate exponent. Then, we have that $L^q(\Omega)$ is the dual space of $L^p(\Omega)$, denoted by $(L^p)' = L^q$. With the duality pairing $\langle u, v \rangle_\Omega := \int_\Omega u(\mathbf{x}) v(\mathbf{x}) \, \mathrm{d}\mathbf{x}$ and for $v \in L^q(\Omega)$, the dual norm follows:*

$$\|v\|_{L^q(\Omega)} = \sup_{u \in L^p(\Omega) \setminus \{0\}} \frac{\langle u, v \rangle_\Omega}{\|u\|_{L^p(\Omega)}}.$$

Theorem A.8 (Hölder's inequality). *Let $1 \le p \le \infty$ and q its conjugate exponent. Assume that $f \in L^p(\Omega)$ and $g \in L^q(\Omega)$. Then, it holds $fg \in L^1(\Omega)$ and*

$$\int_\Omega |fg| \le \|f\|_{L^p(\Omega)} \|g\|_{L^q(\Omega)}.$$

Remark A.9. *For $p = q = 2$, this is called the **Cauchy-Schwarz inequality**:*

$$|(f,g)_{L^2(\Omega)}| \le \|f\|_{L^2(\Omega)} \|g\|_{L^2(\Omega)}, \quad \forall f, g \in L^2(\Omega).$$

Proposition A.10 (General Hölder's inequality). *Let $1 \le p_1, p_2, q \le \infty$ such that $\frac{1}{p_1} + \frac{1}{p_2} = \frac{1}{q}$. Assume that $f \in L^{p_1}(\Omega)$ and $g \in L^{p_2}(\Omega)$. Then, it holds $fg \in L^q(\Omega)$ and*

$$\|fg\|_{L^q(\Omega)} \le \|f\|_{L^{p_1}(\Omega)} \|g\|_{L^{p_2}(\Omega)}.$$

Proposition A.11 (Minkowski inequality). *Let $1 \le p \le \infty$ and assume $f, g \in L^p(\Omega)$. Then, it holds $f + g \in L^p(\Omega)$ and*

$$\|f + g\|_{L^p(\Omega)} \le \|f\|_{L^p(\Omega)} + \|g\|_{L^p(\Omega)}.$$

By $C^0(\Omega)$ we denote the space of all continuous functions in Ω. The support of a function $f \in C^0(\Omega)$ is defined by

$$\operatorname{supp} f := \overline{\{x \in \Omega \mid f(x) \neq 0\}}.$$

The space of all bounded and infinitely differentiable functions on Ω is denoted by $C^\infty(\Omega)$ and we set $C_0^\infty(\Omega)$ to be the space of all functions $f \in C^\infty(\Omega)$ with compact support in Ω. By $C^k(\Omega)$ we denote the space of k-times continuously differentiable functions and $C_0^k(\Omega)$ is the subspace of $C^k(\Omega)$ that contains functions vanishing on the boundary.
Denote $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_d)$, for $\alpha_i \in \mathbb{N}_0$, as a multi-index with absolute value $|\boldsymbol{\alpha}| = \sum_{i=1}^d \alpha_i$ and faculty $\boldsymbol{\alpha}! = \prod_{i=1}^d \alpha_i!$. For $\mathbf{x} \in \mathbb{R}^d$ it holds $\mathbf{x}^{\boldsymbol{\alpha}} = \prod_{i=1}^d x_i^{\alpha_i}$ and for a sufficiently smooth function u, we write the partial derivative of order $|\boldsymbol{\alpha}|$ as:

$$D^{\boldsymbol{\alpha}} u(\mathbf{x}) = \frac{\partial^{|\boldsymbol{\alpha}|}}{\prod_{i=1}^d \partial x_i^{\alpha_i}} u(\mathbf{x}).$$

Definition A.12 (Weak derivative). *Let $f \in L^1(\Omega)$ and $\boldsymbol{\alpha} \in \mathbb{N}_0^d$. Then, $g \in L^1(\Omega)$ is called the weak derivative of order $\boldsymbol{\alpha}$ of f, denoted by $g = D^{\boldsymbol{\alpha}} f$, if for all $\phi \in C_0^\infty(\Omega)$ it holds*

$$\int_\Omega f D^{\boldsymbol{\alpha}} \phi = (-1)^{|\boldsymbol{\alpha}|} \int_\Omega g\phi.$$

Remark A.13. *If f has a continuous partial derivative $D^{\boldsymbol{\alpha}} f$ in the classical sense, then $D^{\boldsymbol{\alpha}} f$ is also a weak derivative.*

Definition A.14 ($W^{m,p}$ spaces). *For $m \in \mathbb{N}$ and $1 \le p \le \infty$ we define*

$$W^{m,p}(\Omega) := \{u \in L^p(\Omega) \mid D^{\boldsymbol{\alpha}} u \in L^p(\Omega) \; \forall \, |\boldsymbol{\alpha}| \le m\},$$

equipped with the norm

$$\|u\|_{W^{m,p}(\Omega)} := \left\{ \sum_{|\boldsymbol{\alpha}| \le m} \|D^{\boldsymbol{\alpha}} u\|_{L^p(\Omega)}^p \right\}^{1/p}$$

and the semi-norm

$$|u|_{W^{m,p}(\Omega)} := \left\{ \sum_{|\boldsymbol{\alpha}| = m} \|D^{\boldsymbol{\alpha}} u\|_{L^p(\Omega)}^p \right\}^{1/p},$$

for $1 \le p < \infty$. For $p = \infty$, the norm is

$$\|u\|_{W^{m,\infty}(\Omega)} := \max_{|\boldsymbol{\alpha}| \le m} \|D^{\boldsymbol{\alpha}} u\|_{L^\infty(\Omega)}.$$

We define the completion of $C_0^\infty(\Omega)$ with respect to the norm $\|\cdot\|_{W^{m,p}(\Omega)}$ as $W_0^{m,p}(\Omega)$.

Theorem A.15.

a) $W^{m,p}$ *is a Banach space.*

b) For $m = 0$*, it holds that* $W^{0,p} = L^p$ *and if* $1 \le p < \infty$*, we have further* $W_0^{0,p} = L^p$*.*

c) The space $H^m := W^{m,2}$ *is a Hilbert space with scalar product*

$$(u,v)_{H^m(\Omega)} := \sum_{|\boldsymbol{\alpha}| \le m} (D^{\boldsymbol{\alpha}} u, D^{\boldsymbol{\alpha}} v)_{L^2(\Omega)} \quad \text{for } u, v \in H^m(\Omega).$$

Theorem A.16. *Let* $1 \le p < \infty$*,* q *its conjugate exponent and* $m \ge 1$*. The dual space of* $W_0^{m,p}$ *is* $(W_0^{m,p})' = W^{-m,q}$*. In particular,* $\left(H_0^1(\Omega)\right)' = H^{-1}(\Omega)$*.*

In order to state the important Sobolev embedding theorems, we first have to recall some definitions:

Definition A.17 (Embedding). *Let* X *and* Y *be Banach spaces. We call* X *continuously embedded in* Y*, denoted by* $X \hookrightarrow Y$*, if* $X \subset Y$ *and if the inclusion* $\iota : X \to Y$ *is continuous, i.e.,* $\exists C$ *such that*

$$\|x\|_Y \le C \|x\|_X \quad \forall x \in X.$$

We call X *compactly embedded in* Y*, denoted by* $X \overset{c}{\hookrightarrow} Y$*, if* $X \subset Y$ *and if the inclusion* $\iota : X \to Y$ *is compact.*

Theorem A.18. *Suppose* Ω *is bounded and* $1 \le p \le q \le \infty$*. Then, it holds*

$$L^q(\Omega) \hookrightarrow L^p(\Omega).$$

Theorem A.19.

a) It holds:

$$H_0^1(\Omega) \hookrightarrow L^2(\Omega) \hookrightarrow H^{-1}(\Omega),$$

b) For any m *it further holds:*

$$W_0^{m,p}(\Omega) \hookrightarrow W^{m,p}(\Omega) \hookrightarrow L^p(\Omega).$$

Theorem A.20. *Let* $\Omega \subset \mathbb{R}^d$ *be a bounded domain with Lipschitz boundary. Let* $k, m \in \mathbb{N}_0$ *and* $1 \le p, q < \infty$*. Then, it holds:*

a) If $k - \frac{d}{p} \ge m - \frac{d}{q}$ *and* $k \ge m$*, then it holds*

$$W^{k,p}(\Omega) \hookrightarrow W^{m,q}(\Omega).$$

b) If $k - \frac{d}{p} > m - \frac{d}{q}$ *and* $k > m$*, then it holds*

$$W^{k,p}(\Omega) \overset{c}{\hookrightarrow} W^{m,q}(\Omega).$$

c) For any open, bounded subset $\Omega \subset \mathbb{R}^d$*, the above statements hold for the spaces* $W_0^{k,p}(\Omega)$*.*

Theorem A.21 (Sobolev embedding). *Let* $\Omega \subset \mathbb{R}^d$ *be a domain with Lipschitz boundary. Let* $j, m \in \mathbb{N}_0$ *and* $1 \le p < \infty$*.*

a) If $mp < d$*, then it holds*

$$W^{j+m,p}(\Omega) \hookrightarrow W^{j,q}(\Omega), \quad p \le q \le \frac{dp}{d - mp}$$

and in particular

$$W^{m,p}(\Omega) \hookrightarrow L^q(\Omega), \quad p \le q \le \frac{dp}{d - mp}.$$

b) If $mp = d$, then it holds

$$W^{j+m,p}(\Omega) \hookrightarrow W^{j,q}(\Omega), \quad p \le q < \infty, \tag{A.6}$$

so in particular

$$W^{m,p}(\Omega) \hookrightarrow L^q(\Omega), \quad p \le q < \infty. \tag{A.7}$$

If $p = 1$, i.e. $m = d$, then (A.6) and (A.7) also hold for $q = \infty$.

c) If $mp > d$, then it holds

$$W^{j+m,p}(\Omega) \hookrightarrow W^{j,q}(\Omega), \quad p \le q \le \infty, \quad m \ge 1$$

and in particular

$$W^{m,p}(\Omega) \hookrightarrow L^q(\Omega), \quad p \le q \le \infty.$$

Moreover, it holds

$$W^{j+m,p}(\Omega) \overset{c}{\hookrightarrow} C^j(\overline{\Omega}), \quad m \ge 1.$$

d) All the embeddings are valid for arbitrary bounded domains, provided that $W^{m,p}$ is replaced by $W_0^{m,p}$.

For boundary value problems, we are interested in the value of a function on the boundary. Since the Sobolev spaces are actually equivalence classes, i.e., two functions are identified if they differ only on a set of measure zero, and because the boundary Γ is a set of measure zero, we can not define the value on the boundary via a simple restriction. Therefore, we need to define a special operator.

Definition A.22 (Trace operator). *For $1 \le p < \infty$, define the map*

$$\gamma : W^{1,p}(\Omega) \to L^p(\Gamma).$$

Then, the trace of $u \in W^{1,p}(\Omega)$ is defined by $\gamma(u)$ and also denoted by $u|_\Gamma$.

Theorem A.23 (Trace theorem). *Let Ω be a bounded domain with Lipschitz boundary Γ and let $1 \le p < \infty$.*

a) If $u \in W^{1,p}(\Omega)$, then in fact $\gamma(u) = u|_\Gamma \in W^{1-(1/p),p}(\Gamma)$ and

$$\|\gamma(u)\|_{W^{1-(1/p),p}(\Gamma)} \le C\|u\|_{W^{1,p}(\Omega)}.$$

b) Furthermore, the trace operator γ is surjective from $W^{1,p}(\Omega)$ onto $W^{1-(1/p),p}(\Gamma)$.

c) The kernel of the trace operator is $W_0^{1,p}(\Omega)$, i.e.,

$$W_0^{1,p}(\Omega) = \left\{u \in W^{1,p}(\Omega) \mid \gamma(u) = 0\right\}.$$

We did not define the Sobolev spaces $W^{s,p}$ of fractional order here, for a definition and further details see [3].
Now we are able to use boundary values, which we understand in the sense of traces. This means that a boundary condition $u = g$ on Γ is understood as $\gamma(u) = g$ on Γ.
For the a posteriori error estimation, the following definition is important.

Definition A.24. *We define the space*

$$H(\Omega, \operatorname{div}) := \left\{\mathbf{q} \in L^2(\Omega) \mid \operatorname{div} \mathbf{q} \in L^2(\Omega)\right\}.$$

Theorem A.25. *$H(\Omega, \operatorname{div})$ is a Hilbert space equipped with the scalar product*

$$(\mathbf{p}, \mathbf{q})_{\operatorname{div}} := \int_\Omega \left(\langle \mathbf{p}, \mathbf{q}\rangle + \operatorname{div} \mathbf{p} \operatorname{div} \mathbf{q}\right).$$

The corresponding norm is induced by the scalar product.

Remark A.26. *Note that it holds:*

$$H^1(\Omega) \subset H(\Omega, \operatorname{div}) \subset L^2(\Omega).$$

This is clear, since for $\mathbf{u} \in H^1(\Omega)$ *it holds* $\nabla \mathbf{u} \in L^2(\Omega)$ *by definition and* $\operatorname{div} \mathbf{u}$ *is the sum of the diagonal of* $\nabla \mathbf{u}$*, hence it follows* $\operatorname{div} \mathbf{u} \in L^2(\Omega)$*. Conversely, for* $\mathbf{u} \in H(\Omega, \operatorname{div})$*, it does not hold* $\mathbf{u} \in H^1(\Omega)$*, since we do not know whether* $\partial_{x_1} u_2$ *and* $\partial_{x_2} u_1$ *are in* $L^2(\Omega)$*.*

Definition A.27. *For functions* $\mathbf{q}$ *in* $L^2(\Omega)$ *and a symmetric and positive definite matrix* $\mathbf{A}(\mathbf{x}) \in \mathbb{R}^{d\times d}$ *for* $\mathbf{x} \in \Omega$*, we denote the following energy and complementary energy norms*

$$\|\mathbf{q}\|_{\mathbf{A}}^2 := \int_\Omega \langle \mathbf{A}\mathbf{q}, \mathbf{q}\rangle \,\mathrm{d}x \quad \text{and} \quad \|\mathbf{q}\|_{\mathbf{A}^{-1}}^2 := \int_\Omega \langle \mathbf{A}^{-1}\mathbf{q}, \mathbf{q}\rangle \,\mathrm{d}x.$$

A.3 Poincaré and Friedrichs Inequality

In the following, we consider as before a bounded domain $\Omega \subset \mathbb{R}^d$ with Lipschitz boundary. The theorems and inequalities, if not indicated differently, are taken from [26] and [10].

Theorem A.28 (Friedrichs inequality). *For* $u \in H_0^1(\Omega)$*, the following inequality*

$$\|u\|_{L^2(\Omega)} \le C_{F\Omega} \|\nabla u\|_{L^2(\Omega)}$$

holds with a constant $C_{F\Omega}$ *independent of* u*. If* $u \in H^1(\Omega)$*, the inequality takes a more general form:*

$$\|u\|_{L^2(\Omega)}^2 \le C_{F\Omega}^2 \left(\|\nabla u\|_{L^2(\Omega)}^2 + \int_\Gamma |u|^2 \,\mathrm{d}s \right).$$

Theorem A.29 (Poincaré inequality). *For* $u \in H^1(\Omega)$*, the following inequality*

$$\|u\|_{L^2(\Omega)}^2 \le C_{P\Omega} \left(\left(\int_\Omega u \,\mathrm{d}\mathbf{x} \right)^2 + \|\nabla u\|_{L^2(\Omega)}^2 \right)$$

holds with a constant $C_{P\Omega}$ *independent of* u*. If* u *fulfils* $\langle u \rangle_\Omega = 0$*, with the mean value* $\langle \cdot \rangle_\Omega$ *defined in Definition 2.9, then it holds*

$$\|u\|_{L^2(\Omega)} \le C_{P\Omega} \|\nabla u\|_{L^2(\Omega)}.$$

Theorem A.30 (Poincaré-Wirtinger inequality). *For* $u \in W^{1,p}(\Omega)$*, the following inequality*

$$\|u - \langle u \rangle_\Omega\|_{L^2(\Omega)} \le C_{PW} \|\nabla u\|_{L^2(\Omega)}$$

holds with a constant C_{PW} *independent of* u*, where* $\langle \cdot \rangle_\Omega$ *is defined in Definition 2.9.*

Theorem A.31 (Trace inequality). *For* $u \in H^1(\Omega)$*, the following inequality*

$$\|\gamma(u)\|_{L^2(\Gamma)} \le C_{T\Gamma} \|\nabla u\|_{L^2(\Omega)}$$

holds with a constant $C_{T\Gamma}$ *independent of* u*.*

The Friedrichs and Poincaré constant will be used in a posteriori error majorants, therefore we need upper bounds for them in order to have fully computable error estimates. For convex domains, the Poincaré constant can be estimated as

$$C_{P\Omega} \le \frac{\operatorname{diam} \Omega}{\pi}, \tag{A.8}$$

for a proof see [25]. If Ω is included in a rectangle $R := \left\{ \mathbf{x} \in \mathbb{R}^2 \mid a_i < x_i < b_i,\ b_i - a_i = l_i,\ i = 1, 2 \right\}$, we have the following estimate for the Friedrichs constant:

$$C_{F\Omega} \le \frac{l_1 l_2}{\pi \sqrt{l_1^2 + l_2^2}}. \tag{A.9}$$

In our examples we usually consider $\Omega = (0,1)^2$, hence we have the following upper bounds:

$$C_{P\Omega} \leq \frac{\sqrt{2}}{\pi}, \quad C_{F\Omega} \leq \frac{1}{\pi\sqrt{2}}. \tag{A.10}$$

With the following approach we get an upper bound by considering eigenvalues: In the first Friedrichs inequality of Theorem A.28, we can see that

$$\frac{1}{C_{F\Omega}} \leq \frac{\|\nabla u\|_{L^2(\Omega)}}{\|u\|_{L^2(\Omega)}}.$$

For a finite element basis $\{\psi_i\}_{i=1}^N$, we have the stiffness matrix $S = (\int_\Omega \langle \nabla\psi_j, \nabla\psi_i \rangle)_{i,j=1}^N$. It follows that for a finite element approximation u_h of u, it holds:

$$\|\nabla u_h\|^2_{L^2(\Omega)} = u_h^\top S u_h.$$

If we use this identity and plug into the above inequality the eigenvector v_h of S, corresponding to the eigenvalue λ_h, then we get

$$\frac{\|\nabla v_h\|^2_{L^2(\Omega)}}{\|v_h\|^2_{L^2(\Omega)}} = \frac{v_h^\top S v_h}{\|v_h\|^2_{L^2(\Omega)}} = \lambda_h.$$

Therefore, lower bounds of the minimal eigenvalue $\lambda_{\min}$ give an upper bound for the Friedrichs constant:

$$\frac{1}{C^2_{F\Omega}} = \lambda_{\min} := \inf_{v \in H^1_0(\Omega)\setminus\{0\}} \frac{\|\nabla v\|^2_{L^2(\Omega)}}{\|v\|^2_{L^2(\Omega)}}.$$

A.4 Clément Operator

Consider the Clément interpolation operator $\mathbf{C}_h : L^2(\Omega) \to V_h \subset H^1(\Omega)$, e.g. see [14]. V_h denotes the finite element space with basis functions ψ_i as defined in (2.32). For a function $u \in L^2(\Omega)$, $\mathbf{C}_h$ is defined by

$$\mathbf{C}_h u := \sum_{i=1}^N \gamma_i(u)\psi_i.$$

In this definition, $\gamma_i : C^\infty(\Omega) \to \mathbb{R}$ is a functional associated to ψ_i and to the patch of the node $\mathbf{x}_i$ corresponding to ψ_i, described in detail in [14]. In [9], the Clément interpolation is explained as an operator $\mathbf{C}_h : H^1(\Omega) \to V_h$, where $\gamma_i : L^2(\omega_i) \to \mathcal{P}_0$ is a local L^2-projection onto the constant functions, with ω_i the patch of $\mathbf{x}_i$. In [16] the interpolation operator $\mathbf{C}_h$ is again defined by a local L^2-projection, but additionaly onto macroelements consisting of element patches. Their definition is applicable for functions in $L^1(\Omega)$. All three definitions have the drawback, that they do not fit boundary conditions. This issue has been considered in [32], but not for functions in L^2.
From [16] we have the following stability and approximation conditions of the Clément interpolation operator:

Proposition A.32.

a) Let $1 \leq p < +\infty$ and $0 \leq m \leq 1$. Then, there exists a constant c such that

$$\|\mathbf{C}_h v\|_{W^{m,p}(\Omega)} \leq c\|v\|_{W^{m,p}(\Omega)} \quad \forall h, \forall v \in W^{m,p}(\Omega).$$

b) Let $1 \leq p < +\infty$ and $0 \leq m \leq l \leq t+1$, where t such that $\psi_i \in \mathcal{P}_t$. Then, there exists a constant c such that

$$\|v - \mathbf{C}_h v\|_{W^{m,p}(T)} \leq c h_T^{l-m} \|v\|_{W^{l,p}(\omega_T)} \quad \forall h, \forall T \in \mathcal{T}_j, \forall v \in W^{l,p}(\omega_T).$$

For $h = \max_{T \in \mathcal{T}_j} h_T$, the last property can easily be extended to norms on Ω. We will in particular use the following inequalities, which follow directly from Proposition A.32:

Proposition A.33.

a) For all $v \in L^2(\Omega)$, there exist constants $C_{L^2 \leftarrow L^2}$ and c such that

$$\begin{aligned} \|\mathbf{C}_h v\|_{L^2(\Omega)} &\leq C_{L^2 \leftarrow L^2} \|v\|_{L^2(\Omega)}, \\ \|v - \mathbf{C}_h v\|_{L^2(\Omega)} &\leq c \|v\|_{L^2(\Omega)}. \end{aligned}$$

b) For all $v \in H^1(\Omega)$, there exist constants $C_{H^1 \leftarrow H^1}$, c_1 and c_2 such that

$$\begin{aligned} \|\mathbf{C}_h v\|_{H^1(\Omega)} &\leq C_{H^1 \leftarrow H^1} \|v\|_{H^1(\Omega)}, \\ \|v - \mathbf{C}_h v\|_{L^2(\Omega)} &\leq c_1 h \|v\|_{H^1(\Omega)}, \\ \|v - \mathbf{C}_h v\|_{H^1(\Omega)} &\leq c_2 \|v\|_{H^1(\Omega)}. \end{aligned}$$

Bibliography

[1] A. Abdulle. On A Priori Error Analysis of Fully Discrete Heterogeneous Multiscale FEM. *Multiscale Modeling & Simulation*, 4(2):447–459, 2005.

[2] A. Abdulle. A Priori and a Posteriori Error Analysis for Numerical Homogenization: a Unified Framework. *Series in Contemporary Applied Mathematics*, 16:280–305, 2011.

[3] R. A. Adams. *Sobolev Spaces*. Academic Press, New York, 1975.

[4] H. W. Alt. *Lineare Funktionalanalysis. Eine anwendungsorientierte Einführung*. Springer, Berlin, 2006.

[5] R. E. Bank and J. Xu. Asymptotically Exact a Posteriori Error Estimators, Part I: Grids with Superconvergence. *SIAM Journal on Numerical Analysis*, 41(6):2294–2312, 2003.

[6] R. E. Bank and J. Xu. Asymptotically Exact a Posteriori Error Estimators, Part II: General Unstructured Grids. *SIAM Journal on Numerical Analysis*, 41(6):2313–2332, 2004.

[7] S. Bartels and P. Schreier. *Local Coarsening of Triangulations Created by Bisections*. Univ., SFB 611, 2010.

[8] A. Bensoussan, J.-L. Lions, and G. Papanicolau. *Asymptotic Analysis for Periodic Structures*. North-Holland, Amsterdam, New York, Oxford, 1978.

[9] D. Braess. *Finite Elemente. Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*. Springer, Berlin, 2007.

[10] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer Science & Business Media, 2010.

[11] L. Chen and C. Zhang. A Coarsening Algorithm on Adaptive Grids by Newest Vertex Bisection and its Applications. *Journal of Computational Mathematics*, 28(6):767–789, 2010.

[12] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems*. North-Holland, Amsterdam, New York, Oxford, 1978.

[13] D. Cioranescu and P. Donato. *An Introduction to Homogenization*, volume 17. Oxford Lecture Series in Mathematics and its Applications, Oxford University Press, 1999.

[14] P. Clément. Approximation by Finite Element Functions Using Local Regularization. *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique*, 9(2):77–84, 1975.

[15] C. F. Dunkl. oral communication, 2016.

[16] A. Ern and J.-L. Guermond. *Theory and Practice of Finite Elements*, volume 159. Springer Science & Business Media, 2004.

[17] S. Funken, D. Praetorius, and P. Wissgott. Efficient Implementation of Adaptive P1-FEM in MATLAB. *Computational Methods in Applied Mathematics*, 11(4):460–490, 2011.

[18] W. Gautschi. *Numerical Analysis*. Springer Science & Business Media, 2012.

[19] D. Gilbarg and N. S. Trudinger. *Elliptic Partial Differential Equations of Second Order*, volume 224. Springer Science & Business Media, 2001.

[20] P. Grisvard. *Elliptic Problems in Nonsmooth Domains*, volume 69. SIAM, 1985.

[21] W. Hackbusch. *Theorie und Numerik elliptischer Differentialgleichungen*. Springer, 1986.

[22] V. V. Jikov, S. M. Kozlov, and O. A. Oleinik. *Homogenization of Differential Operators and Integral Functionals*. Springer, Berlin, 1994.

[23] J. N. Lyness and R. Cools. A survey of numerical cubature over triangles. In *Proceedings of Symposia in Applied Mathematics*, volume 48, pages 127–150, 1994.

[24] O. Mali, P. Neittaanmäki, and S. Repin. *Accuracy Verification Methods: Theory and Algorithms*, volume 32. Springer Science & Business Media, 2013.

[25] L. E. Payne and H. F. Weinberger. An Optimal Poincaré Inequality for Convex Domains. *Archive for Rational Mechanics and Analysis*, 5(1):286–292, 1960.

[26] S. Repin. *A Posteriori Estimates for Partial Differential Equations*, volume 4. Walter de Gruyter, Berlin, 2008.

[27] S. Repin, T. Samrowski, and S. Sauter. Combined a Posteriori Modeling-Discretization Error Estimate for Elliptic Problems with Complicated Interfaces. *ESAIM: Mathematical Modelling and Numerical Analysis*, 46(6):1389–1405, 2012.

[28] S. Repin, T. Samrowski, and S. Sauter. Estimates of the Modeling Error Generated by Homogenization of an Elliptic Boundary Value Problem. *Journal of Numerical Mathematics*, 24(1):1–15, 2016.

[29] S. Repin, S. Sauter, and A. Smolianski. A Posteriori Error Estimation for the Dirichlet Problem with Account of the Error in the Approximation of Boundary Conditions. *Computing*, 70(3):205–233, 2003.

[30] R. Schnyder. Time and Space Adaptive Solution to Retarded Potential Integral Equations. Master's thesis, University of Zurich, 2013.

[31] C. Schwab and A.-M. Matache. Generalized FEM for Homogenization Problems. *Multiscale and Multiresolution Methods. Lecture Notes in Computational Science and Engineering*, 20:197–237, 2002.

[32] L. R. Scott and S. Zhang. Finite Element Interpolation of Nonsmooth Functions Satisfying Boundary Conditions. *Mathematics of Computation*, 54(190):483–493, 1990.

[33] O. Steinbach. *Numerical Approximation Methods for Elliptic Boundary Value Problems: Finite and Boundary Elements*. Springer Science & Business Media, 2008.

[34] L. B. Wahlbin. *Superconvergence in Galerkin Finite Element Methods*, volume 1605. Lecture Notes in Mathematics, Springer, Berlin, 1995.

Curriculum Vitae

Name	MEIER-ROHR
Vorname	Stephanie
Geburtsdatum	21. November 1988
Heimatort	Staufen AG

Ausbildung

08/2004 - 12/2007	**Gymnasium Liestal, Matura** Schwerpunkt: Anwendungen der Mathematik/Physik
09/2008 - 09/2011	**Universität Basel, Bachelor of Science in Mathematics**
09/2011 - 09/2013	**Universität Basel, Master of Science in Mathematics** Vertiefungsrichtungen: Numerik und Algebra Masterarbeit: Sparse grid quadrature in high dimensions
01/2014 - heute	Anstellung als **Doktorandin** an der **Universität Zürich** Vertiefungsrichtung: Numerische Analysis